République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

ECOLE NATIONALE POLYTECHNIQUE

DEPARTEMENT D'HYDRAULIQUE

Projet de fin d'étude

Présenté par :

STIHI Abdelkader Hichem

Pour l'obtention du diplôme d'Ingénieur d'Etat

En HYDRAULIQUE

Thème

Approche de l'évaluation de la sujétion de service public induite par la mobilisation et le transfert de l'eau de surface des barrages en exploitation

Proposé et dirigé par : Dr A. BERMAD

Promotion juin 2012

ملخّص

يهدف هذا العمل إلى إيجاد طريقة لمحاكاة تكلفة المتر المكعب من مياه كلّ سدّ، وبالتالي تقدير قيمة إخضاع القطاع العامّ المخصصة للوكالة الوطنيّة للسّدود و التّحويلات حتّى تنجز مهمّتها المتمثّلة في تسيير و استغلال السّدود. ساهم تحليل المكونات الرئيسية في خطوة أولى في إبراز العلاقة الموجودة بين المتغيرات المميّزة للسّدود وإظهار بعض السّدود الخاصّة. وأنشئت عدة نماذج مختلفة من حيث المتغيرات التفسيرية التي تتضمّنها عن طريق تقنيّة الانحدار المتعدد. تتراوح النتائج المحقّقة من ممتازة إلى جيدة بنسبة مقبولة من الدقة في تقدير متوسط تكلفة الوحدة من مياه السّدود

الكلمات المفتاحية : السّدود, تكلفة ,محاكاة .

Résumé

Ce travail a pour but d'établir une démarche permettant de simuler le coût de revient unitaire de l'eau par barrage, et par conséquent évaluer le montant de la sujétion de service public alloué à l'ANBT dans sa mission d'exploitation des barrages. Une analyse en composantes principales a permis dans un premier temps de mettre à jour les relations existant entre les variables caractérisant les barrages et de distinguer les barrages atypiques. Plusieurs modèles, comportant différentes combinaisons de variables explicatives, ont été établis par la suite à l'aide de la méthode de régression multiple. Les résultats obtenus varient d'excellents à satisfaisants et présentent une bonne précision dans l'estimation du coût de revient unitaire moyen.

Mots clés: Simulation, Coût, Barrages.

Abstract

This work aims to establish an approach to simulate the unit cost of water for each dam, and therefore assess the amount of the subjection of public service allocated to ANBT in its mission of dam exploitation. A principal component analysis resulted in a first step that shows the relationship between the variables characterizing dams and distinguished atypical ones. Several models with different combinations of explanatory variables were established later using the multiple regression method. The results vary from excellent to good and have a satisfactory accuracy in estimating the average unit cost.

Key words: Simulating, Cost, Dams.

Remerciements

Je remercie mes parents, mes frères et ma sœur pour leur soutien inconditionnel tout au long de ces années,

Je remercie mon promoteur, Mr Bermad pour son suivi et son aide dans l'élaboration de ce travail.

Je tiens aussi à remercier Mr Benyoucef de bien vouloir présider le jury, et messieurs les jurés de bien vouloir juger la qualité de mon travail.

Mes remerciements vont à Mr Haglaouane, pour sa disponibilité et pour avoir mis à ma disposition les données nécessaires à cette étude.

Merci à tous les enseignants qui ont participé à ma formation, et à qui je dois en grande partie mon futur statut d'ingénieur en hydraulique.

*Un grand merci à tous mes camarades et amis, qui m'ont permis de passer ces trois dernières années moins péniblement que prévu, et plus particulièrement Rami, Karim et Fella qui sont là depuis le début de l'aventure^^*

Sans oublier l'équipe d'hydraulique, Djamel, Zaki, Youcef, Said, Chouaib, Karim, Abdelali, Yasmine, Wissem, Yacine…et tous les autres (qui sont trop nombreux pour être tous cités) avec qui j'ai passé de très bons moments.

Enfin, je remercie tous ceux qui ont contribué de près ou de loin à l'aboutissement de ce projet.

Table des matières

Liste des tableaux

Liste des figures

Introduction générale

Les besoins croissants en eau ont nécessité la redéfinition du système de mobilisation des eaux superficielles. Outre l'importance de l'investissement dans la construction et la maintenance des barrages, nous assistons aujourd'hui à des mutations profondes du secteur, notamment les aspects liés à la gestion de l'eau ainsi que l'émergence des concepts liés au marché.

Depuis 2000, de nombreuses réformes ont été engagées, notamment la création du Ministère des Ressources en Eau (MRE), la restructuration et la transformation des agences et le fort accroissement des programmes d'investissement.

L'Agence Nationale des Barrages et Transferts, occupe une place particulière dans ce dispositif de maîtrise de la chaîne de l'eau, en tant que premier outil opérationnel du Ministère, axé sur la mobilisation des ressources en eaux superficielles.

Afin d'assurer son autonomie de fonctionnement, l'Agence Nationale des Barrages et Transferts doit disposer des fonds nécessaires à cet effet.

Son caractère d'Entreprise à caractère Industriel et Commercial devrait lui permettre de subvenir à ses besoins via la vente de l'eau brute aux entreprises de distribution, or, le coût de cette vente n'est pas soumis aux variations du marché mais, pour des raisons externes à l'Agence, est administré par l'état à hauteur de 1 DA/m^3 .

Ce revenu, ne reflétant pas la réalité du marché, est insuffisant pour couvrir l'ensemble des dépenses encourues par l'Agence, et c'est pour cette raison, qu'une subvention de l'état est indispensable, c'est la sujétion de service public.

Problématique

Le prix de vente de l'eau brute de l'ANBT aux utilisateurs est fixé à 1 DA/m^3.

Les rentrées résultant de cette vente sont loin de couvrir toutes les charges que l'Agence doit supporter dans sa mission d'exploitation des barrages.

Il faut par conséquent, déterminer la différence entre les dépenses totales et les revenus de l'Agence, c'est cette différence qui devra être prise en charge par la sujétion de service public.

Ce travail a pour but d'établir une démarche permettant de déterminer le coût de revient réel de l'eau à l'avenir, à défaut de pouvoir évaluer la sujétion de service public faute de disponibilité de données détaillées et actualisées.

Après cette introduction mettant en relief l'importance de la sujétion de service public dans le bon fonctionnement de l'ANBT, ainsi que la présentation de la problématique à résoudre, nous allons suivre le plan décrit ci-dessous :

- Un premier chapitre consacré à la présentation de l'ANBT, de ses activités, et une introduction au domaine de la comptabilité analytique nécessaire à la bonne compréhension du déroulement de ce travail.
- Le second chapitre abordera les aspects théoriques de l'Analyse en Composantes Principales,
- Le troisième quant à lui sera consacré à la méthode de la régression multiple qui sera la méthode de modélisation adoptée,
- Enfin, le quatrième et dernier chapitre constituera la partie application de cette étude, les résultats obtenus y seront présentés et critiqués.

CHAPITRE I

COMPTABILITE ANALYTIQUE

I.1 Introduction

Ce chapitre a pour but de présenter l'Agence Nationale des Barrages et Transferts, ses activités, et la place qu'elle occupe dans la chaîne de l'eau. Il permettra également de familiariser le lecteur avec certaines notions de base liées au domaine de la comptabilité analytique, notamment les concepts de coût qui seront utilisés ultérieurement dans l'évaluation de la sujétion de service public.

I.2 Présentation de l'ANBT

L'Agence Nationale des Barrages par abréviation « ANB» est créée par le décret n°85.163 du 11 janvier 1985 avec un statut d'E.P.A (Entreprise Publique à caractère Administratif).

Dans le cadre de la transformation de la forme juridique des établissements sous tutelle du Ministère des Ressources en Eau en vertu du 2eme article du décret n°101-05 du 23 mars 2005, le statut de l'Agence Nationale des Barrages, est réaménagé dans sa nature juridique en E.P.I.C (Etablissement Public à caractère Industriel et Commercial) dénommé « Agence Nationale des Barrages et Transferts » par abréviation « ANBT».

L'Agence Nationale des Barrages et Transferts en sa qualité de maitre d'ouvrage délégué, sous la tutelle du Ministère des Ressources en Eau est chargée de la mise en œuvre des programmes nationaux de mobilisation et de transfert des eaux superficielles. Elle est également chargée de la gestion de l'exploitation de l'ensemble de ces ouvrages de mobilisation et de transfert qui, à l'horizon 2025, atteindront une capacité globale de 9 milliards de mètres cubes.

Les articles 07,09 et 10 du nouveau statut, stipulent que l'ANBT a pour mission principale, la production et la fourniture d'eau aux établissements et aux régies communales chargées de sa distribution, d'assurer la prise en charge des activités de gestion, d'exploitation et de maintenance des ouvrages en exploitation, dans le cadre de la mobilisation et du transfert des ressources en eaux superficielles.

L'objectif de la transformation en EPIC, vise à donner à l'ANBT une plus grande autonomie d'organisation et de gestion pour mieux la responsabiliser sur ses missions et les résultats qui en sont attendus. Cette mutation lui permettra d'acquérir une relative indépendance vis-à-vis du budget de l'Etat par l'existence de revenus propres issus de :

- la facturation de l'eau brute ou traitée qu'elle délivrera aux organismes utilisateurs,

- la rémunération pour sa mission de maîtrise d'ouvrage et d'œuvre déléguée, appelée à se réduire de plus en plus.

L'ANBT doit assurer en permanence les travaux d'entretien de plus de 60 barrages et transferts en exploitation.

Pour faire face aux dépenses d'entretien et de maintenance qui ne sont plus imputées directement sur le budget de l'Etat, l'ANBT doit disposer en permanence de ressources nécessaires. C'est à ce titre que l'ANBT demande l'attribution de la dotation de sujétions de service public conformément à l'article n° 24 du décret exécutif n° 101-05 du 23/05/2005.

I.3 Sujétion de service public

I.3.1 Définition

La sujétion de service public est une rémunération, par l'état, d'un organisme public, couvrant l'ensemble des charges induites par les tâches confiées à cet organisme conformément à un cahier de charges prédéfini.

I.3.2 Activités de L'ANBT

La sujétion de service public, dans le cadre de l'ANBT découle de la mission « exploitation des ouvrages » qui lui est confiée et qui s'articule autour de trois activités principales :

I.3.2.1 Exploitation de la ressource en eau

C'est à l'ANBT que revient la charge de préserver la ressource en eau et d'en assurer la mise à disposition aux différents utilisateurs.

Outre les opérations liées à la fourniture de l'eau brute, l'ANBT est chargée de :

- Suivre la qualité des eaux ;
- Protéger la ressource ;
- Suivre les apports et l'état des réserves en eau ;
- Suivre l'état de l'envasement des retenues ;
- Collecter, traiter et conserver l'ensemble des données hydrométéorologiques ;
- Effectuer les chasses des sédiments ;
- Prendre en charge la gestion des crues ;
- Assurer la gestion des évènements exceptionnels ;
- Suivre les programmes de soutirage périodiques et procéder à la répartition des ressources avec les organismes concernés ;

I.3.2.2 Auscultation et surveillance des ouvrages

L'aspect sécuritaire lié aux barrages constitue un enjeu important. Il concerne en premier lieu le barrage lui-même et sa sécurité intrinsèque. La surveillance des barrages est la mission la plus sensible dans toute l'activité de l'agence. Les responsabilités qui en découlent sont considérables.

Les exploitants exécutent quotidiennement, un certain nombre de tâches pour permettre un suivi du comportement de l'ouvrage, tel que :

- Procéder au contrôle et à la vérification des différentes mesures effectuées sur les appareillages d'auscultation des barrages avec mise en graphique des résultats de mesures et suivi de l'évolution des différents mouvements de l'ouvrage et de sa fondation ;
- Assurer l'auscultation permanente des ouvrages par un suivi des différents mouvements subis et du comportement de leurs fondations ;
- Interpréter les mesures internes et externes d'auscultation en rapport avec les données techniques et scientifiques en vigueur et en fonction des données de l'exploitation ;
- Elaborer les plans de sécurité (ORSEC) en collaboration avec les autorités locales et les structures concernées ;
- Assurer des inspections visuelles périodiques des ouvrages ;
- Elaborer des rapports de comportement annuels.

I.3.2.3 Entretien des ouvrages

Un entretien périodique et une bonne maintenance, sont les seuls garants de la longévité et de la pérennité du fonctionnement des barrages.

Cette opération relève des priorités de l'exploitation. Il ne suffit pas de rénover les équipements usés ou défaillants, il est aussi nécessaire de les entretenir et de les conserver.

Ces travaux ont comme seul objet de réduire la probabilité de défaillance ou de dégradation des installations. Les activités correspondantes sont enclenchées selon :

- Un échéancier établi à partir d'un nombre prédéterminé d'unités d'usage ;
- Des critères prédéterminés significatifs de l'état de dégradation de l'équipement.

Dans ce cadre, l'exploitant exécute périodiquement les tâches suivantes :

- Elaboration des programmes de maintenance préventive ;
- Entretien des équipements de pompage et de drainage des eaux ;
- Entretien et remplacement des joints d'étanchéité des vannes de servitude ;
- Entretien des équipements de levage et de manutention ;
- Reprise des revêtements extérieurs des conduites ;
- Entretien et réparation des routes et des accès ;
- Entretien des talus et bermes ;
- Entretien des galeries et des ouvrages de prise ;
- Entretien des équipements électriques ;
- Etc.

I.4 Comptabilité analytique

I.4.1 Définition

La comptabilité analytique est un instrument à usage interne pour la gestion des sous-ensembles distingués dans l'activité de l'entreprise.

Elle se distingue techniquement par le recensement des charges par « destination ». Il s'agit de déterminer quelle part d'une charge peut être attribuée à tel produit ou telle activité, à tel sous ensemble de l'entreprise. La comptabilité analytique permet ainsi de calculer divers types de coûts dont les usages sont multiples.

I.4.2 Historique

La connaissance des coûts est, depuis le début de la révolution industrielle, un impératif de base de toute prise de décision. La notion de comptabilité industrielle manifesta très tôt cette nécessité et se traduisit par la mise en place par les comptables, mais aussi par les ingénieurs et les techniciens, de systèmes de calcul aptes à les aider dans leur gestion. Les historiens recensent ainsi des systèmes précurseurs de comptabilité industrielle dès la fin du XVIIe siècle en Grande Bretagne (notamment dans les forges et fonderies de la région de Sheffield). En France une doctrine apparaît à partir des années 1860 et on situe en 1885 l'apparition du premier manuel.

Appelée d'abord comptabilité industrielle puis comptabilité analytique d'exploitation, la comptabilité de gestion désigne l'ensemble des éléments du système comptable considérés du point de vue de l'intérêt qu'ils présentent pour la gestion interne.

Entre 1947 et 1999, les rédacteurs du plan comptable général français ont voulu normaliser la comptabilité analytique au même titre que la comptabilité financière (ou comptabilité générale). Depuis 1999, la comptabilité de gestion n'est plus normalisée. Ses méthodes et son organisation doivent être adaptées aux particularités et aux besoins spécifiques de chaque entreprise ou organisation. De plus, l'objectif de la normalisation est de faciliter les comparaisons interentreprises. Or, cet objectif ne concerne pas la comptabilité de gestion qui est à usage interne et dont les résultats sont rarement divulgués.

I.4.3 Principaux concepts de coût

I.4.3.1 Objet de coût

Un objet de coût est tout élément pour lequel une mesure séparée du coût est jugée utile.

Exemples : un produit (l'eau), un service (stockage de l'eau), un réseau de distribution, un département (siège administratif)…etc

I.4.3.2 Inducteur de coût

Un inducteur de coût se définit comme étant un quelconque facteur susceptible d'avoir un impact sur le coût d'un objet de coût. Autrement dit, toute modification de l'inducteur de coût entraîne un changement dans le coût total de l'objet de coût. C'est donc un facteur qui cause les coûts.

Exemple : Nombre d'unités produites, nombre de barrages…etc

I.4.3.3 Coût de revient

Le coût de revient d'un produit représente le prix d'une unité de ce produit qui, outre le coût des produits pris en stock, inclut une part de charges « hors production » telles que des charges de recherche et développement et des charges d'administration.

Dans cette étude, le coût de revient de l'eau représente le prix d'une unité d'eau distribuée par l'ANBT, toutes charges comprises, d'une installation hydraulique : le barrage.

On distingue deux types de charges :

I.4.3.3.1 Les charges directes

Une charge directe est une charge dont il est facilement observable qu'elle a été encourue pour un objet de coût spécifique et peut donc être affectée, sans aucune ambiguïté, à cet objet de coût.

Ces charges sont affectées directement au coût d'un produit ou d'un service car la consommation de ces charges pour chaque type de produit ou service est connue.

I.4.3.3.2 Les charges indirectes

Une charge indirecte est une charge qui n'est pas associée spécifiquement et uniquement à un objet de coût car il est impossible de l'affecter directement à un objet de coût particulier. Ce sont les charges qui sont consommées par plusieurs objets de coût.

Celles-ci ne sont intégrées dans le calcul des coûts qu'après des calculs permettant de définir la partie de ces charges relative à chaque objet de coût. Ces calculs sont réalisés à l'aide de clés de répartition conventionnelles (au prorata des capacités, des effectifs, du nombre d'unités produites…etc), les résultats sont finalement imputés aux objets considérés lors de l'analyse.

Il importe de faire une distinction : le caractère direct ou indirect d'une charge dépend de l'objet de coût considéré.

I.4.3.4 Cas d'étude

Dans cette étude, et afin de déterminer les coûts unitaires, il nous faut au préalable répertorier les charges relatives à l'exploitation de chaque barrage, celles-ci sont les suivantes :

I.4.3.4.1 Charges Directes

Les besoins prévisionnels prennent en charge l'ensemble des taches afférentes à l'activité de gestion des barrages et des infrastructures annexes.

L'Agence Nationale des barrages et Transferts qui gère plus de 60 ouvrages de mobilisation, exécute d'une manière quotidienne les opérations d'entretien et de maintenance de l'ensemble des ouvrages et équipements composant ce parc.

Il est à signaler que sur chaque barrage, l'ANBT dispose de locaux administratifs et d'accueil (chambre de passage), des magasins et ateliers, des chambres de commande, des tours de prise et des chambres de vannes ainsi que des galeries qui atteignent des longueurs d'une dizaine de Kilomètres.

Les équipements installés sur les ouvrages sont de différents types :

- Vannes ;
- Tuyauteries ;
- Centrales hydrauliques ;
- Armoires de commande ;
- Stations de pompage ;
- Appareils d'auscultation ;
- Ascenseurs et monte-charges ;
- Ponts roulants et moyens de manutention ;

  Chaque barrage, est doté de moyens de locomotion et de liaison. Ces moyens nécessitent un entretien permanent et continu.

D'un autre coté, Les barrages sont soumis en permanence aux aléas climatiques et à l'usure due au fonctionnement continu des équipements, en particulier ceux immergés qui sont en constance agressés par les eaux. Leur maintenance est vitale.

La vulnérabilité des versants, est souvent à l'origine d'éboulis et de glissements de terrain qui nécessitent des interventions très lourdes pour le dégagement des accès et des ouvrages d'entrée.

Les barrages s'étendent sur des superficies très élargies et dans des régions éloignées dont le relief est très accidenté. Cette contrainte oblige les exploitants à procéder à l'entretien régulier de ces espaces.

Les charges directes sont liées aux produits de consommation et à l'outillage nécessaires pour l'accomplissement de ces tâches d'entretien et de maintenance courante. Une grande partie de ces charges provient de la consommation électrique.

I.4.3.4.2 Charges indirectes

Cette partie des dépenses, provient essentiellement des frais induits par le soutien technique et logistique qu'assure la direction générale au profit des exploitants, à titre d'exemple :

- Elaboration de la paie ;
- Affectation des budgets ;
- Consolidation des bilans comptables ;
- Consolidation du suivi de la ressource ;

- Interprétation des mesures d'auscultation ;
- Expertises ponctuelles ;
- Etc.

A ces charges il faut ajouter les dépenses induites par le siège de l'ANBT et les différentes unités d'exploitation, qui représentent une grande partie des charges indirectes totales.

Les charges indirectes seront imputées aux barrages lors du calcul du coût unitaire par barrage.

I.4.3.4.3 Charges classées par nature

Afin de mieux cerner la justification des coûts nécessaires à une bonne gestion des barrages en exploitation, voici, en détail, les différentes charges classées et regroupées par nature, et destinées à une comptabilité dite « générale » qui ne sera toutefois pas abordée lors de cette étude :

I.4.3.4.3.1 Produits d'entretien

- Ciments ordinaires à prise rapide ;
- Agrégats de différentes granulométries / Résines ordinaires et spéciales pour l'étanchéité ;
- Agrégats pour mortiers et bétons ;
- Peintures ordinaires pour bâtiments et locaux ;
- Peintures industrielles pour la protection des équipements métalliques ;
- Produits de droguerie tels les acides, décapants, produits chimiques…etc .

I.4.3.4.3.2 Charges annexes

- Paiement des consommations d'eau et d'électricité ;
- Paiement des redevances téléphoniques, fax, télex. ;
- Acquisition et développement documentation, films vidéo, matériel de reproduction et de projection ;
- Assurance matériel roulant ;
- Matériel de sécurité et de surveillance et de réparation ;

I.4.3.4.3.3 Fournitures diverses et consommables

- Equipements électriques, bobines, câbles, fusibles, lampes, etc ;
- Produits de quincaillerie ;
- Tuyauteries et outillages ;
- Matériaux de menuiserie et machines outil portatives ;
- Gaz liquéfiés et équipement de soudure ;
- Habillement, produits pharmaceutiques ;
- Pièces de rechange pour les engins ;
- Renouvellement des équipements de mesure et de contrôle ;
- Acquisition des appareils de mesure ;
- Acquisition des équipements informatiques ;
- Carburants et lubrifiants pour véhicules.

I.4.3.4.3.4 Prestation et service d'entretien des équipements et ouvrages

- Travaux de réparation des équipements hydromécaniques ;
- Travaux de réaménagement des locaux, ouvrages et réseaux, canaux d'assainissement ;
- Travaux de génie civil des ouvrages ;
- Réparation de matériel roulant ;
- Réparation des équipements de liaison téléphonique et de communication ;
- Installation et réparation des équipements électriques.

I.4.3.4.3.5 Charges du personnel

- Salaires ;
- Frais de déplacement et de mission ;
- Frais de formation.

I.4.3.4.3.6 Assistance et expertises spécialisées

I.4.3.5 Coût standard

Le coût de revient cité précédemment est un coût réel calculé à posteriori, on parle aussi de coût constaté, dans le but d'analyser l'activité passée. Mais il est également nécessaire de prévoir l'avenir et de fixer des objectifs. C'est le rôle des **coûts standards** également appelés coûts préétablis. Ils permettent :

- De contrôler les conditions d'exploitation en analysant les écarts entre coût standard et coût réel.
- De servir de base pour l'élaboration des devis.

I.5 Conclusion

Ce chapitre nous a permis d'introduire les principales définitions et notions utilisées dans la comptabilité analytique, telles que les coûts et les charges. Ces dernières ont été répertoriées afin de justifier le coût de revient de l'eau résultant de la mobilisation de la ressource à travers les différents barrages en exploitation.

CHAPITRE II

ANALYSE EN COMPOSANTES PRINCIPALES

II.1 Introduction

L'Analyse en Composantes principales (ACP) fait partie du groupe des méthodes descriptives multidimensionnelles appelées méthodes factorielles. Ces méthodes, dans la mesure où ce sont des méthodes descriptives, ne s'appuient pas sur un modèle probabiliste, mais dépendent d'un modèle géométrique. L'ACP propose, à partir d'un tableau rectangulaire de données comportant les valeurs de p variables quantitatives pour n unités (appelées aussi individus), des représentations géométriques de ces unités et de ces variables. Ces données peuvent être issues d'une procédure d''echantillonnage ou bien de l'observation d'une population toute entière. Les représentations des unités permettent de voir s'il existe une structure, non connue à priori, sur cet ensemble d'unités. De façon analogue, les représentations des variables permettent d'étudier les structures de liaisons linéaires sur l'ensemble des variables considérées. Ainsi, on cherchera si l'on peut distinguer des groupes dans l'ensemble des unités en regardant quelles sont les unités qui se ressemblent, celles qui se distinguent des autres, etc. Pour les variables, on cherchera quelles sont celles qui sont très corrélées entre elles, celles qui, au contraire ne sont pas corrélées aux autres, etc.

Enfin, comme pour toute méthode descriptive, réaliser une ACP n'est pas une fin en soi. L'ACP servira à mieux connaître les données sur lesquelles on travaille, à détecter éventuellement des valeurs suspectes…etc. On pourra aussi, a posteriori, se servir des représentations fournies par l'ACP pour illustrer certains résultats utiles à une nouvelle approche telles que la régression ou la classification automatique.

II.2 Historique

Bien que l'étude de la structure de vastes ensembles de données soit récente, les principes dont les méthodes d'analyse de données s'inspirent sont anciens.

En ce qui concerne l'analyse factorielle, il faut remonter aux travaux de Ch. Spearman (1904) qui introduit pour la première fois le concept de facteur ; il cherche, derrière les notes obtenues par de nombreux sujets à de nombreux tests, une variable explicative cachée : le facteur général d'aptitude (analyse factorielle au sens des psychologues).

C'est vers les années 30 que se pose le problème de la recherche de plusieurs facteurs (travaux de C. Burt et de L.L Thurstone) ; on cherche deux puis plusieurs facteurs : mémoire, intelligence, etc. ''non observables directement mais susceptibles d'expliquer au sens statistique du terme les nombreuses notes obtenues par les sujets''. Comme on le constate il s'agissait déjà de résumer à l'aide d'un petit nombre de facteurs une information multidimensionnelle.

Puis, l'analyse factorielle en composantes principales est développée par H. Hotelling (1933), mais dont on peut faire remonter le principe à K. Pearson (1901) : ''les individus colonnes du tableau à analyser étant considérés comme des vecteurs d'un espace à p dimensions, on proposait de réduire la dimension de l'espace en projetant le nuage des points individus sur le sous-espace de dimension *q* (*q* petit fixé) permettant d'ajuster le mieux le nuage. ''.(*Ambapour, 2003*)

II.3 Principe de la méthode

Le principe général de l'A.C.P. est de réduire la dimension des données initiales (qui est *p* si l'on considère *p* variables quantitatives), en remplaçant les *p* variables initiales par *q* facteurs appropriés ($q < p$).

Les données, toujours centrées, doivent en plus être réduites lorsque les variables sont hétérogènes. Les *q* facteurs cherchés sont des moyennes pondérées des variables initiales. Leur choix se fait en maximisant la dispersion des individus selon ces facteurs (autrement dit, les facteurs retenus doivent être de variance maximum).

Plus simplement, L'idée à la base de l'analyse en composantes principales est de pouvoir expliquer ou rendre compte de la variance observée dans la masse de données initiales en se limitant à un nombre réduit de composantes, définies comme étant des transformations mathématiques pures et simples des variables initiales.

II.4 Notion D'individu et de variable

Dans une série de données on distingue généralement deux ensembles : les individus et les variables qui sont les caractères relatifs à ces individus. Le terme "individu" peut designer selon les cas : une année d'observation, un barrage…L'ensemble des individus peut provenir d'un échantillonnage dans une population ou il peut s'agir de la population toute entière (Comme c'est le cas pour nous).

- L'individu "i" est décrit par le vecteur appartenant à R^n :

$$X_i = \{X_{ij} / j = 1 \text{ à } n\} \quad \text{..... (II-1)}$$

Sur un individu, on relève un certain nombre de caractères (dits aussi variables) désignant en général un paramètre intervenant dans le phénomène à étudier

- Le caractère (ou variable) "j" est décrit par le vecteur de R^p :

$$X_j = \{X_{ij} / i = 1 \text{ à } p\} \quad \text{..... (II-2)}$$

Le terme X_{ij} est un nombre réel qui représente la mesure de la variable X_j sur l'individu i.

Si l'ensemble des individus doit être homogène, l'ensemble des variables peut être hétérogène.

II.4.1 Les données à analyser

Les données se présentent sous forme d'un tableau rectangulaire dont les colonnes sont des variables quantitatives notées *Xi (j=1…p)* et les lignes les n individus sur lesquels ces variables sont observées *(i=1…n).* Les observations du tableau seront donc les *Xij* correspondant respectivement au *ième* individu et à la *jème* variable.

$$[X]=\begin{vmatrix} X_{11} & X_{12} & . & . & X_{1j} & . & . & X_{1P} \\ X_{21} & X_{22} & . & . & X_{2j} & . & . & X_{2P} \\ . & . & . & . & . & . & . & . \\ X_{i1} & X_{i2} & . & . & X_{ij} & . & . & X_{iP} \\ . & . & . & . & . & . & . & . \\ X_{N1} & X_{N2} & . & . & X_{Nj} & . & . & X_{NP} \end{vmatrix} \quad \text{…… (II-3)}$$

Il faut noter que le nombre *p* de variables doit être égal au moins à 2 et que le nombre d'individus au moins égal à *p*. Généralement on dispose d'une dizaine de variables et de centaines (voire milliers) d'individus.

II.5 Définition algébrique

Pour chaque variable, nous pouvons évaluer sa moyenne $\overline{X_j}$ et son écart type S_j :

Moyenne : La moyenne est une mesure statistique caractérisant les éléments d'un ensemble de quantités : elle exprime la grandeur qu'aurait chacun des membres de l'ensemble s'ils étaient tous identiques sans changer la dimension globale de l'ensemble. Dans notre cas, la moyenne est une moyenne arithmétique définie par

$$\overline{X_j}=\frac{1}{n}\sum_{i=1}^{n}X_{ij} \quad \text{……..…… (II-4)}$$

Ecart-type : C'est une mesure de la dispersion d'une variable aléatoire réelle autour de sa moyenne. Il est défini comme la racine carrée de la variance.

$$S_j=\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_{ij}-\overline{X_j}\right)^2\right]^{1/2} \quad \text{…….. (II-5)}$$

Le coefficient de covariance entre les variables X_j et X_k est donné par :

$$Cov(X_j, X_k) = \frac{1}{n}\sum_{i=1}^{n}\left(X_{ij} - \overline{X_j}\right)*\left(X_{ik} - \overline{X_{ik}}\right) \quad \text{..... (II-6)}$$

Le coefficient de corrélation entre les variables X_j et X_k est donné par :

$$Cor(X_j, X_k) = \frac{Cov(X_j, X_k)}{S_j * S_k} = \frac{\sum_{i=1}^{n}\left(X_{ij} - \overline{X_j}\right)*\left(X_{ik} - \overline{X_k}\right)}{\left[\sum_{i=1}^{n}\left(X_{ij} - \overline{X_j}\right)^2 * \sum_{i=1}^{n}\left(X_{ik} - \overline{X_k}\right)^2\right]^{1/2}} \quad \text{.....(II-7)}$$

Le tableau [X] est remplacé par un tableau [Y] (individus x nouvelles variables) en réduisant le nombre de variables nécessaires pour décrire les individus, avec une perte minimale d'informations, ces nouvelles variables sont appelées composantes principales (ou CP).
Calculer les composantes principales notées C_j revient à déterminer N relations linéaires entre les variables X_j :

$$\begin{cases} C_1 = a_{10} + a_{11}X_1 + ... + a_{1j}X_j + ... + a_{1N}X_N \\ C_2 = a_{20} + a_{21}X_1 + ... + a_{2j}X_j + ... + a_{2N}X_N \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ C_j = a_{j0} + a_{j1}X_1 + ... + a_{jj}X_j + ... + a_{jN}X_N \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ C_N = a_{N0} + a_{N1}X_1 + ... + a_{Nj}X_j + ... + a_{NN}X_N \end{cases} \quad \text{.....(II-8)}$$

Où C_j : $j^{ème}$ Composante Principale ;

X_j : Vecteur variable initiale ;

a_{jl} : Coefficient du système.

Notons au passage que les termes a_{j0} désignent le vecteur permettant la translation de l'origine de l'ancien repère vers le centre de gravité du nuage de points. Un centrage des données initiales annule les coefficients a_{j0}. (*Hamriche et al. 1993*)

II.6 Définition géométrique

L'Analyse en Composantes Principales est puissante par son support géométrique, la méthode consiste à rechercher un premier axe qui soit le plus proche possible de tous les points au sens des moindres carrés, tel que la somme des carrés des distances des *n* points à cet axe serait minimale, ou encore la projection de ces derniers sur cet axe possède une dispersion maximale. Cet axe est appelé "axe factoriel".

Un second axe est obtenu après projection des *n* points sur un hyperplan orthogonal au premier axe, tel que la dispersion des projections des *n* points sur celui-ci serait toujours maximale, et le procédé se réitère *p* fois.

Nous obtenons ainsi un nouveau système d'axes défini par les nouvelles variables dites composantes principales. (*Hamriche et al. 1993)*

II.7 Formulation mathématique du problème

La recherche des composantes principales est faite sous deux contraintes :

- Elles doivent être indépendantes, c'est à dire, prises deux à deux, elles présentent obligatoirement des corrélations nulles.
- Les axes factoriels doivent être déterminés par ordre d'importance décroissante, le premier axe expliquera le maximum de la variance totale tandis que le second expliquera le maximum de la variance résiduelle non expliquée par le premier, jusqu'au dernier axe. Mais l'expérience a montré qu'un nombre Q d'axes nettement inférieur à P suffit pour donner le maximum d'informations.

II.7.1 Formulation matricielle

L'objectif de l'ACP étant de maximiser la variance, la variance étant définie comme la moyenne des carrés des écarts par rapport à la moyenne, la formulation matricielle du problème est la suivante :

Soient les matrices colonnes V_1, V_2,...,V_q de dimension (Nx1) représentant l'hyperplan formé par les axes principaux vérifiant les conditions de normalité et d'orthogonalité :

$$\begin{cases} \vec{V}_i \cdot \vec{V}_i = 1 \\ \vec{V}_i \cdot \vec{V}_j = 0 \end{cases} \rightarrow \begin{cases} i = 1 \rightarrow Q \\ j = 1 \rightarrow Q \end{cases} \quad \text{.....} \quad \text{(II-9)}$$

On veut maximiser la quantité : $\sum_{j=1}^{Q} Var(C_j)$ …… (II-10)

Sachant que : $\sum_{j=1}^{Q} Var(C_j) = V_j^t \cdot [M] \cdot [R] \cdot [M] \cdot V_j$ …..……..…… (II-11)

C_j : Composante principale d'ordre j ;

$[R]$: Matrice des covariances des variables (X_1, X_2, ... X_p) ;

$[M]$: Métrique définissant le produit scalaire sur l'espace R^P.

II.7.2 Choix de la métrique

La métrique [M] possède deux options classiques :

* $[M] = 1$: Matrice identité.

La covariance sera utilisée afin de quantifier les relations inter variables, on parlera alors d'une *ACP canonique.*

* $[M] = D^2_{1/\sigma}$ …….. (II-12)

Tel que :

$$D_{1/\sigma^2} = \begin{vmatrix} 1/\sigma_1^2 & 0 & & & & 0 \\ 0 & 1/\sigma_2^2 & 0 & & & \\ & 0 & . & & & \\ & & 0 & 1/\sigma_i^2 & 0 & \\ & & & & . & \\ & & & & . & 0 \\ 0 & & & & 0 & 1/\sigma_p^2 \end{vmatrix}$$ …..…..…… (II-13)

On utilise généralement cette métrique pour pallier au problème de l'hétérogénéité des caractères (variables), et éviter l'influence du choix d'unité des variables. Dans ce cas on parlera d'ACP normée, elle est équivalente à une ACP canonique effectuée sur des variables centrées réduites.

Les données ainsi transformées se présentent sous forme d'une matrice dont toutes les variables sont de moyenne nulle et d'écart type unité.

II.8 Procédé d'application de l'ACP

II.8.1 Calcul de la matrice de covariance

La matrice des covariances, notée $[R]$ est la base de l'ACP, elle est obtenue en appliquant la relation suivante :

$$R = (1/n)\ X^{t}.\ M\ .\ X \quad \text{..... (II-14)}$$

$[R]$: Matrice de covariance de dimension (p*p) ;

$[X]$: Matrice de données ;

$[X]^t$: Matrice transposée de $[X]$;

$[M]$: Métrique.

II.8.2 Recherche des axes principaux

Le but est de construire un nouveau système d'axes avec un minimum de variables assurant un maximum de variance.

II.8.2.1 Recherche du premier axe

Telle que mentionnée précédemment, la contribution maximale est donnée par le premier axe principal, nous devons donc maximiser la variance relative à celui-ci. La recherche du premier axe principal consiste à résoudre le problème

$$\begin{cases} MaxVar(C_1) \\ V_1^t.[M].V_1 = 1 \end{cases} \quad \text{..... (II-15)}$$

Nous pouvons exprimer la variance de C_1 à l'aide de la matrice des covariances $[R]$ du vecteur aléatoire $X = (X_1, X_2, ...X_j, ...X_p)$:

$$Var(C_1) = V_1^t.[M].[R].[M].V_1 \quad \text{..... (II-16)}$$

En utilisant la méthode du multiplicateur de LAGRANGE nous pouvons écrire :

$$L = V_1^t[M].[R].[M].V_1 - \lambda_1(V_1^t.[M].V_1 - 1) \quad \text{..... (II-17)}$$

La dérivée par rapport à V est nécessairement nulle :

$$\partial L / \partial V_1 = 2.[M].[R].[M].V_1 - 2\lambda_1.[M].V_1 = 0 \quad \text{..... (II-18)}$$

Puisque la matrice $[M]$ est inversible :

$$[R].[M].V_1 = \lambda_1.V_1 \quad \text{..... (II-19)}$$

Donc V_1 est le vecteur propre de la matrice $[R]\cdot[M]$. Il suffit de choisir comme vecteur V_1 le vecteur propre associé à la plus grande valeur propre λ_1 de la matrice $[R]\cdot[M]$ pour maximiser la variance de C_1.

II.8.2.2 Recherche du second axe

Nous cherchons à déterminer le vecteur unitaire V_2 tel que la composante C_2 soit de variance maximale et non corrélée à C_1.

Sachant que :

$$Var(C_2) = V_2^t.[M].[R].[M].V_2 \quad \text{..... (II-20)}$$

$$V_2^t.[M].V_2 = 1$$

$$COV(C_1, C_2) = 0$$

L'expression de $COV(C_1, C_2)$ est donnée par :

$$COV(C_1, C_2) = V_1^t.[M].[R].[M].V_2 \quad \text{..... (II-21)}$$

Comme la covariance ne tient pas compte de l'ordre nous pouvons écrire :

$$COV(C_1,C_2) = COV(C_2,C_1) = V_2^t[M][R][M]V_1 = 0 \quad \text{..... (II-21)}$$

Or nous savons que V_1 est un vecteur propre de $[R]\cdot[M]$ associé à la valeur propre λ_1. Nous en déduisons que :

$$COV(C_1, C_2) = \lambda_1 V_2^t.V_1 = 0 \quad \text{..... (II-22)}$$

Une covariance nulle entre C_1 et C_2 est équivalente à l'orthogonalité des vecteurs V_1 et V_2 :

$$COV(C_1,C_2) = 0 \Leftrightarrow V_1.V_2 = 0 \quad \text{..... (II-23)}$$

En appliquant la même méthode pour la recherche du deuxième axe, nous aurons :

$$L=V_2^t.[M][R][M]V_2-\lambda_2(V_2^t.[M]V_2-1)-\mu_2(V_2^t.[M]V_1) \quad \text{..... (II-24)}$$

$$\partial L/\partial V_2=2.[M][R][M]V_2-2\lambda_2.[M]V_2-\mu_2.[M]V_1=0 \quad \text{..... (II-25)}$$

En simplifiant par [M] nous obtenons :

$$2.[R][M]V_2-2\lambda_2.V_2-\mu_2.V_1=0 \quad \text{..... (II-26)}$$

Nous multiplions à gauche par $V_1^t.[M]$ l'équation (III-22) s'écrit alors :

$$2.V_1^t.[M][R][M]V_2-2\lambda_2.V_1^t.[M]V_2-\mu_2.V_1^t.[M]V_1=0 \quad \text{..... (II-27)}$$

Or (par hypothèse) $V_1^t \cdot [M] \cdot V_2 = 0$ ..... (II-28)

Donc :

$$V_1^t \cdot [M] \cdot [R] \cdot V_2 = V_2^t \cdot [M] \cdot [R] \cdot [M] \cdot V_1 = \lambda_1 \cdot V_2^t \cdot [M] \cdot V_1 = 0 \quad \text{........ (II-29)}$$

Puisque le vecteur V_1 est unitaire:

$$V_1^t.[M]V_1=1 \quad \text{..... (II-30)}$$

Le multiplicateur de Lagrange μ_2 est donc nul, et nous sommes ramenés au problème précédent. Nous pouvons donc énoncer la définition suivante : le second axe est défini par le vecteur V_2, vecteur propre unitaire de la matrice $[R] \cdot [M]$ orthogonal à V_1 et associé à la plus grande valeur propre λ_2 inférieure ou égale à λ_1.

II.8.2.3 Recherche des autres axes

En itérant le procédé, nous déterminons donc les valeurs propres et les vecteurs propres de la matrice $[R] \cdot [M]$ pour obtenir la $l^{ère}$ composante principale C_l.

Le vecteur propre unitaire de la matrice $[R] \cdot [M]$ définit le l^{er} axe orthogonal à $(V_1, V_2, ..., V_{l-1})$ et associé à la $l^{ère}$ plus grande valeur propre λ_1. Nous constatons que la mise en équation de ces règles aboutit aux résultats suivants :

- Nous appelons l[er] vecteur principal : le vecteur propre unitaire V_l de la matrice $[R]\cdot[M]$ associée, qui fournit les coefficients qui pondèrent les variables initiales pour le calcul des composantes principales.
- Nous appelons l[er] axe principal, la droite engendrée par le l[er] vecteur principal.
- Chaque composante C_k est portée par le k[ème] axe principal.
- La dispersion des projections des variables sur la composante C_k est mesurée par la valeur propre λ_k .
- Les valeurs sont rangées par ordre décroissant : $\lambda_1 > \lambda_2 > \lambda_3 > \cdots > \lambda_Q$.
- Les CP sont rangées de 1 à Q dans l'ordre des valeurs propres.
- La moyenne de chaque CP est nulle.
- Si nous voulons normer les CP, c'est à dire, imposer à chacune d'entre elles d'avoir un écart type unité il suffit de diviser chacune d'elles par la racine de la variance expliquée $(\lambda_k)^{1/2}$ correspondante.

II.8.3 Calcul des composantes principales

On désigne par CP, la projection du nuage de points initiale sur le nouveau système d'axes fournit par les vecteurs propres. Notons [C] la matrice des CP.

$$[C] = [A]^t [X] \quad \text{..... (II-31)}$$

$[X]$: Matrice des données initiales ;

$[A]^t$: Matrice transposée de la matrice composée par les vecteurs propres.

II.8.4 Représentation graphique

Une fois les résultats numériques obtenus nous passons à la représentation graphique :

II.8.4.1 Variables

- Nous pouvons représenter chaque variable par un point dont les coordonnées sont les covariances avec les CP réduites.

$$Cov(X_j, C_k) = (\lambda)^{1/2} * V_k(j) \quad \text{..... (II-32)}$$

Si nous considérons une variable centrée réduite Y_j au lieu de la variable initiale X_j, les coordonnées de cette nouvelle variable dans le plan factoriel sont calculées par :

$$COV(Y_j, C_i) = COV\left(\frac{X_j}{S_j}, C_i\right) = Cor(X_j, C_i) \quad \text{.....(II-33)}$$

$C_l^{'}$: Composante principale normée ;

S_j : Ecart type de la $j^{ème}$ variable.

Comme la variance de Y_j est égale à 1, cela signifie que son point représentatif se trouve sur une sphère de rayon égal à 1. C'est pourquoi le cercle unité tracé sur chacun des plans étudiés s'appelle cercle de corrélation. La variable sera d'autant mieux expliquée, que son point représentatif se rapproche du cercle et inversement.

II.8.4.2 Individus

Nous pouvons aussi observer la répartition des individus selon les principaux axes choisis. Ainsi deux individus seront proches dans l'espace R^p, s'ils sont proches dans le plan factoriel. Dans le graphique des individus on s'intéresse aux distances inter-individus qu'on peut interpréter comme étant des ressemblances.

En conclusion, nous pouvons dire que l'analyse en composantes principales, dans son aspect qualitatif, permet d'établir deux sortes de bilans :

- Un bilan de liaisons entre les variables pouvant nous renseigner sur les points suivants :
- Quelles sont les variables qui sont liées positivement, et celles qui s'opposent (liées négativement)?
- Existe-t-il une répartition en groupes des variables inter-corrélées ?
- Un bilan de ressemblance entre individus répondant aux questions suivantes :
- Quels sont les individus qui se ressemblent et ceux qui différent ?
- Existe-t-il des groupes homogènes d'individus ?

II.8.4.3 ACP normée

Lors d'une analyse en composantes principales, toutes les variables sont centrées, c'est à dire qu'elles ne sont plus exprimées par rapport à l'origine du nuage de points mais par rapport au centre de gravité de ce nuage. De plus, lorsque les variables ne sont pas toutes exprimées dans la même unité de mesure, l'opération de réduction par l'écart-type permet de rendre comparables les variables puisque toute valeur d'une variable devient un écart à sa moyenne exprimé en nombre d'écarts-types de la variable. Toutes les observations ainsi recodées deviennent des valeurs comparables, de moyenne nulle et d'écart-type unité (ce sont des valeurs adimensionnelles). C'est ce qu'on appelle une analyse en composantes principales normée. (*Morineau*)

Le traitement des variables centrées réduites se fait de la manière suivante :

Soient Z_j les variables obtenues par la transformation ($Z_j = \frac{X_j - \overline{X_j}}{\sigma_{X_j}}$), montrons que la covariance des deux variables centrées réduites (Z_j, Z_k) n'est autre que la corrélation entre les deux variables brutes correspondantes.

Evaluons maintenant les covariances des variables (Z_j, Z_k) :

$$Cov(Z_j, Z_k) = \frac{1}{N}\sum_{i=1}^{N}(Z_{ij} - \overline{Z_j}) * (Z_{ik} - \overline{Z_k})$$

$$= \frac{1}{N}\sum_{i=1}^{N} Z_{ij} * Z_{ik}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_{ij} - \overline{X_k}}{\sigma_{X_j}}\right) * \left(\frac{X_{ik} - \overline{X_k}}{\sigma_{X_k}}\right)$$

$$= \frac{\frac{1}{N}\sum_{i=1}^{N}(X_{ij} - \overline{X_j}) * (X_{ik} - \overline{X_k})}{\sigma_{X_j} * \sigma_{X_k}}$$

$$= \frac{Cov(X_j, X_k)}{\sigma_{X_j} * \sigma_{X_k}}$$

$$= Cor(X_j, X_k) \quad \text{..... (II-34)}$$

II.9 Adéquation de l'ACP

Bien qu'elle soit – à priori – applicable sans hypothèses, l'analyse en composantes principales ne s'adapte pas toujours aux données de l'étude.

Il existe deux critères permettant de mesurer l'adéquation de l'ensemble des données à une analyse en composantes principales :

II.9.1 Test de sphéricité de Bartlett

L'analyse en composantes principales s'accommode assez bien des situations où un certain niveau de multicolinéarité existe entre les données. Cependant, il faut absolument se méfier de la condition dite de « **singularité** » où une variable serait parfaitement corrélée avec une autre

variable ou avec une combinaison de plusieurs variables. Cette condition peut être détectée en calculant le « déterminant» de la matrice de corrélation $/R/$.

Le déterminant est une valeur numérique unique associée à une matrice carrée et qui peut prendre n'importe quelle valeur entre 0.0 et 1.0. Cependant ces deux valeurs extrêmes sont problématiques. En effet, un déterminant de 0.0 indique que la matrice est singulière c'est-à-dire qu'il existe au moins un cas de dépendance linéaire dans la matrice ou, en d'autres mots, qu'une variable peut être entièrement expliquée ou prédite par une combinaison linéaire d'autres variables. Comme le mentionne Field (2000), on ne devrait jamais procéder à une ACP sur une matrice de corrélation dont le déterminant est plus petit que 0.00001 car on considère qu'il y a de très fortes redondances dans les données c.-à-d. qu'elles ne recèlent qu'un seul type d'information. Mathématiquement, les produits de matrices nécessaires à l'estimation ne peuvent être effectués dans une telle situation (il est impossible d'inverser la matrice).

À l'inverse, un déterminant égal à 1.0 correspond lui aussi une condition impropre à l'ACP. Il indique que la matrice de corrélation est une **matrice d'identité.** Le test de Bartlett vise justement à vérifier si l'on s'écarte significativement de cette situation. L'hypothèse nulle du test étant H0 : $/R/ = 1$.

La statistique du test s'écrit :

$$X^2 = -\left(n - 1 - \frac{2p+5}{6}\right) ln\,|R| \quad \text{..... (II-34)}$$

Sous H0, elle suit une loi du χ^2 à $\left(p\,\frac{p-1}{2}\right)$ degrés de liberté.

Pratiquement, si la signification tend vers 0.000, c'est très significatif ; inférieur à 0.05, significatif ; entre 0.05 et 0.10, acceptable et au dessus de 0.10, l'application d'une ACP n'est pas adaptée.

Il faut noter que le test de Bartlett est sensible à la taille de l'échantillon et que lorsque le n est assez grand, les chances de rejeter l'hypothèse nulle sont très élevées. En ce sens, le rejet de l'hypothèse nulle ne garantit pas nécessairement que l'ACP donne de bons résultats; à l'inverse, si le test de Bartlett ne nous permet pas de rejeter l'hypothèse nulle, nous sommes en présence d'une situation vraiment extrême où l'ACP n'est pas justifiable.

II.9.2 Indice KMO (Kaiser-Meyer-Olkin)

Le test de sphéricité de Bartlett mesure si l'ensemble d'une matrice de corrélation possède les propriétés souhaitées pour une ACP. Il est également important d'examiner chacune des variables de façon individuelle et c'est dans cette optique que l'indice KMO appelé aussi MSA (Measure of Sampling Adequacy) est calculé.

Ce test mesure l'importance des coefficients de corrélation par rapport aux coefficients de corrélations partielles.

On sait que les variables sont plus ou moins liées. La corrélation brute entre deux variables est influencée par les (p-2) autres. Nous utilisons la corrélation partielle pour mesurer la relation (nette) entre deux variables en retranchant l'influence des autres. L'indice cherche alors à confronter la corrélation brute avec la corrélation partielle. Si la seconde est nettement plus faible (en valeur absolue), cela veut dire que la liaison est effectivement déterminée par les autres variables. Cela accrédite l'idée de redondance, et donc la possibilité de mettre en place une réduction efficace de l'information.

Le KMO peut être calculé pour la matrice dans sa globalité ou par variable :

II.9.2.1 Indice KMO global

Il est donné par la formule

$$\text{KMO} = \frac{\sum_i \sum_{j \neq i} r_{ij}^2}{\sum_i \sum_{j \neq i} r_{ij}^2 + \sum_i \sum_{j \neq i} a_{ij}^2} \quad \text{..... (II-35)}$$

L'indice KMO varie entre 0 et 1. Un KMO élevé indique qu'il existe une solution factorielle statistiquement acceptable qui représente les relations entre les variables.

Pour être conservée dans une ACP, une variable doit obtenir une mesure KMO dépassant 0.5.

Kaiser (1974) a suggéré une gradation intéressante utilisant les points de référence suivants :

- Inacceptable en dessous de 0,5 ;
- Médiocre entre 0,5 et 0,6 ;
- Moyen entre 0,6 et 0,7 ;
- Bon entre 0,7 et 0,8 ;
- Très bon entre 0,8 et 0,9 ;
- Excellent au dessus de 0,9.

II.9.2.2 Indice KMO par variable

Il est défini par la formule suivante :

$$KMO_j = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} a_{ij}^2} \quad \text{..... (II-36)}$$

Le KMO par variable varie dans le même intervalle que le KMO global, cependant, un KMO_j faible ne veut pas forcément dire que la variable en question doit être rejetée de l'analyse mais qu'elle contribue fortement à la construction d'un axe différent de celui construit par le reste des variables ayant de plus fortes corrélations entre elles.

II.10 Choix du nombre de composantes

Une fois l'adéquation de l'ACP vérifiée, nous devons chercher quel est le nombre optimal de composantes à extraire, pour ce faire, plusieurs méthodes existent, les plus utilisées étant :

II.10.1 Critère de Kaiser (1960)

Dans la matrice de corrélation, les valeurs présentes sur la diagonale principale (des 1), correspondent à la variance de chaque variable.

La variance totale étant don égale au nombre de variables. Lors de l'extraction des composantes principales, cette variance est répartie en maximisant celle de la première composante, la seconde composante expliquera quant à elle une portion de la variance indépendante de la précédente et plus faible qu'elle, et le calcul se poursuit jusqu'à l'extraction de toutes les composantes c'est-à-dire jusqu'à l'explication de toute la variance initiale.

La valeur propre de chaque composante représente la nouvelle variance qu'explique cette composante. C'est sur cette base que Kaiser a établi un critère permettant de déterminer le bon nombre de composantes à extraire.

D'après ce critère, l'extraction doit s'arrêter lorsque l'une des composantes expliquera une variance inférieure à celle expliquée par la variable initiale, autrement dit, dès qu'une valeur propre devient inférieure à 1.

II.10.2 Test d'accumulation de variance (scree test) de Cattell (1966)

En 1966, Cattell a proposé une méthode graphique pour décider du nombre de composantes à extraire. Le test d'accumulation de variance communément appelé « scree test » ou aussi « coude de Cattell » demande que l'on trace un graphique illustrant la taille des valeurs propres des différentes composantes en fonction de leur ordre d'extraction. Le terme « scree » fait référence à un phénomène géomécanique où l'on observe une accumulation de dépôts rocheux au pied d'une montagne, créant ainsi un petit promontoire à l'endroit où le dénivelé de la montagne se transforme brusquement en une pente plus douce. Le critère proposé par Cattell nous amène à arrêter l'extraction des composantes à l'endroit où se manifeste le changement de pente dans le graphique.

La figue ci dessous correspond au test d'accumulation de variance pour nos données fictives. On y constate que la pente change radicalement avec la composante C3. La représentation graphique des variances nous aide à voir que le point C3 appartient beaucoup plus au segment C3 à C7 qu'au segment C1 à C3. Selon le critère de Cattell on devrait donc se limiter à l'extraction des deux premières composantes.

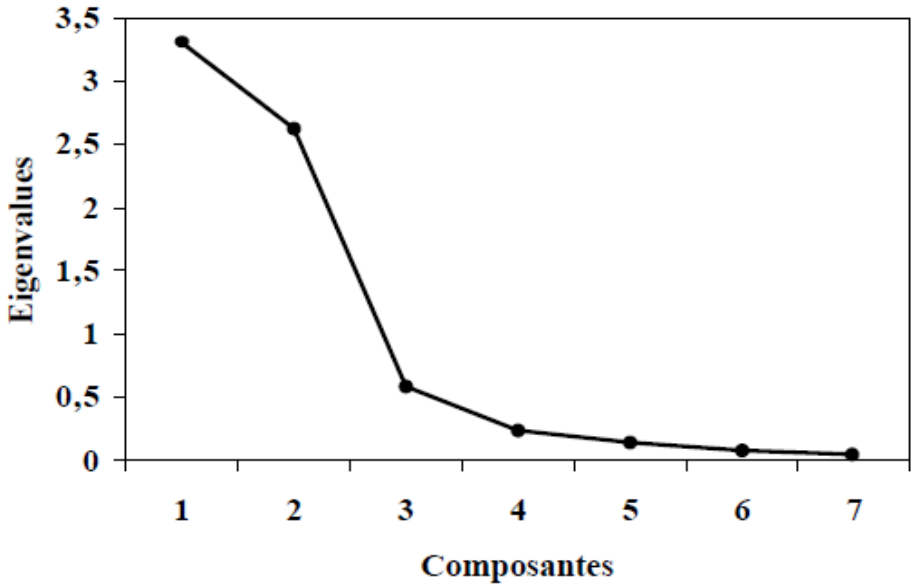

Figure II-1 : Test d'accumulation de variance de Cattell

II.10.3 Analyse parallèle de Horn (1965)

L'approche suggérée par Horn (1965) pour déterminer le nombre de composantes à extraire s'appuie sur un raisonnement très différent des deux précédents. Horn indique qu'il est possible de découvrir par chance une composante pouvant expliquer une certaine proportion de variance, même en partant de données générées complètement au hasard et pour lesquelles aucune dimension réelle n'existe. Cette proportion de variance, expliquée par pure chance, pourrait donc servir comme point de comparaison afin de nous aider à décider si la variance

que nous obtenons dans notre analyse est significativement plus importante que celle observable dans une matrice de données générées de façon aléatoire. L'analyse parallèle consiste donc à mener une ACP sur une matrice de corrélation générée au hasard mais comportant le même nombre de variables et d'individus que notre étude. La série décroissante des valeurs propres calculées sur ces données aléatoires sera alors comparée aux valeurs propres calculées sur les données réelles. Si une composante existe vraiment dans nos données de recherche, sa valeur propre correspondante devrait être significativement plus grande que celle obtenue sur les données aléatoires. Ainsi, Horn recommande de ne conserver pour extraction que les composantes dont les variances sont significativement supérieures à celles obtenues par pure chance. La prise de décision est relativement facilitée si l'on trace un graphique représentant les deux séries de valeurs propres. La figure ci-dessous permet de constater que cette méthode indiquerait deux composantes à extraire. (*Baillargeon, 2003*)

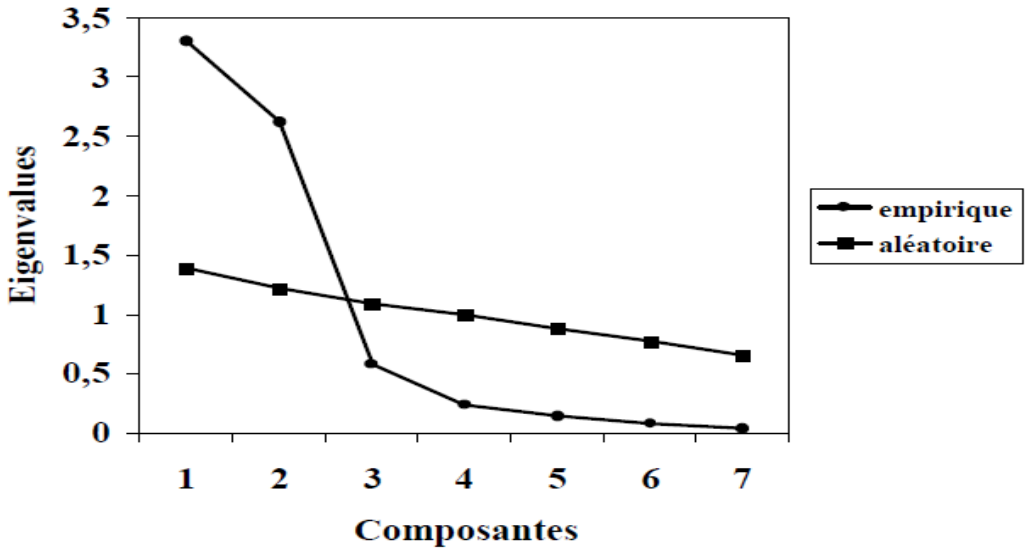

Figure II-2 : Analyse parallèle de Horn

II.11 Conclusion

Nous avons vu dans ce chapitre les différents aspects théoriques de l'Analyse en composantes principales, d'un point de vue analytique et géométrique, la façon de procéder à une telle analyse, les critères permettant d'évaluer la pertinence d'une telle méthode par rapport aux données de l'étude ainsi que les techniques pour déterminer le nombre optimal de composantes à prendre en compte dans l'interprétation des résultats.

Les représentations des variables et des individus dans les plans factoriels définis par les composantes sélectionnées serviront de base à une étude plus approfondie des données telle que la régression que nous verrons dans le prochain chapitre.

CHAPITRE III

REGRESSION MULTIPLE

III.1 Introduction

Il arrive souvent que l'on veuille expliquer la variation d'une variable dépendante par l'action de plusieurs variables explicatives. Lorsque le cas se présente, on peut recourir à la méthode de régression multiple. Dans ce chapitre, nous allons voir les mécanismes et les aspects théoriques relatifs à cette technique.

III.2 Historique

Le terme régression a été introduit par Francis Galton (1822-1911), chercheur britannique du 19éme siècle dans son article « Regression towards mediocrity in hereditary stature » afin de décrire le phénomène biologique qui est que la taille des enfants nés de parents inhabituellement grands ou petits se rapproche de la taille moyenne de la population.

III.3 Forme générale du modèle

La régression linéaire multiple est une généralisation, à P variables explicatives, de la régression linéaire simple.

Etant donné un échantillon ($Yi, Xi1, \ldots, Xip$), $i = 1,\ldots,n$, nous cherchons à expliquer, avec le plus de précision possible, les valeurs prises par Yi, dite variable dépendante, à partir d'une série de variables explicatives (ou indépendantes) $Xi1, \ldots, Xip$. Le modèle théorique, formulé en termes de variables aléatoires, prend la forme

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots b_p X_{ip} + \varepsilon_i.\ i = 1,\ldots, n \ldots\ldots\ldots\ldots \quad \text{(III-1)}$$

où ε_i est l'erreur du modèle qui exprime, ou résume, l'information manquante dans l'explication linéaire des valeurs observées à partir des variables explicatives. Les a_i sont les paramètres à estimer.

L'équation de régression multiple est celle d'un hyperplan à P dimensions qui ne peut pas être représenté concrètement (au-delà de 2 dimensions), les paramètres b_i sont les pentes de cet hyperplan dans les dimensions considérées et sont appelés « coefficients de régression ».[A]

L'hyperplan est ajusté selon le principe des moindres carrés-qui sera traité plus en détail dans ce qui suit.

III.4 Objectifs de la régression multiple

La régression multiple peut être utilisée à plusieurs fins:

- Trouver la meilleure équation linéaire de prévision (modèle) et en évaluer la précision et la signification.
- Estimer la contribution relative de deux ou plusieurs variables explicatives sur la variation d'une variable à expliquer; déceler l'effet complémentaire ou, au contraire, antagoniste entre diverses variables explicatives.
- Juger de l'importance relative de plusieurs variables explicatives sur une variable dépendante en lien avec une théorie causale sous-jacente à la recherche.(*Borcard,2010*)

III.5 Moindres carrés ordinaires

III.5.1 Historique

La paternité des moindres carrés est souvent discutée. Il semble qu'elle ait été découverte indépendamment par Carl Friedrich Gauss en 1795 et Adrien Marie Legendre en 1805. Mais aucun document écrit n'a pu confirmer la date de 1795 qui repose sur les affirmations de C.F.Gauss lui-même. (*Dodge et al. 2004*)

III.5.2 Principe

Lorsque la relation entre Yi et les $X_{i1},\ldots, X_{ip}$ existe, les données mesurées n'appartiennent pas forcément à la droite de régression. Pour tenir compte dans le modèle mathématique des erreurs observées, on considère les données $\{Y_1, Y_2, \ldots, Y_p\}$ comme autant de réalisations d'une variable aléatoire Y et parfois aussi les données $\{(X_{11},\ldots, X_{1p}),\ldots, (X_{p1},\ldots,X_{pp})\}$ comme autant de réalisations de variables aléatoires $X_1, X_2, \ldots, X_p$.

Les données $\{(X_{i1},\ldots, X_{ip}, Y_i), i = 1,\ldots, n\}$ peuvent être assimilées à un nuage de n points. Rechercher une relation affine entre les variables X_i et Y revient à rechercher une droite qui s'ajuste le mieux possible à ce nuage de points. Parmi toutes les droites possibles, on retient celle qui rend minimale la somme des carrés des écarts des valeurs observées Y_i à la droite de régression. Si ε représente cet écart, appelé aussi résidu, Le principe des moindres carrés consiste à rechercher les valeurs des paramètres qui minimisent la somme des carrés des résidus, c'est-à-dire la quantité :

$$E = \sum_{i=0}^{n} \varepsilon_i^2 = \sum_{i=0}^{n}(Y_i - b_0 - b_1 X_{i1} - \ldots - b_p X_{ip})^2 \ldots\ldots\ldots\ldots \text{(III-2)}$$

III.6 Démarche de la régression multiple

Les étapes d'une modélisation par régression multiple sont les suivantes :

- Estimer les paramètres « a_i » en exploitant les données
- Evaluer la précision de ces estimateurs
- Mesurer le pouvoir explicatif global du modèle, en considérant toutes les variables, et de façon individuelle en considérant les variables une à une
- Sélectionner les variables les plus pertinentes
- Evaluer la qualité du modèle lors de la prédiction
- Détecter les observations qui peuvent influencer ou fausser exagérément les résultats (points atypiques) (*Rakotomalala*)

III.6.1 Calcul des coefficients de régression

Le calcul des coefficients de régression peut se faire de plusieurs manières, l'une, utilisée dans les logiciels informatiques repose sur le calcul matriciel. L'autre, présentée ci-dessous se base sur un systèmes de *m* équations à *m* inconnues qui permet dans un premier temps d'obtenir les coefficients de régression centrés et réduits (comme si on faisait une régression sur des variables centrées réduites).

Les valeurs des coefficients de régression pour les valeurs brutes sont ensuite obtenues par multiplication par le rapport des écarts-types de la variable dépendante et de la variable explicative considérée.

Finalement, on calcule la valeur de l'ordonnée à l'origine.

III.6.1.1 Calculs préliminaires

On peut calculer les coefficients de régression et l'ordonnée à l'origine d'une régression multiple en connaissant:

- Les coefficients de corrélation linéaire simple de toutes les paires de variables entre elles (y compris la variable dépendante): r_{x1x2}, r_{x1x3} ... r_{x1y}, ...etc.;
- Les écarts-types de toutes les variables: S_{x1}, S_{x2}, S_{x3}...s_y;
- Les moyennes de toutes les variables.

III.6.1.2 Calcul des coefficients centrés-réduits

On calcule les coefficients centrés-réduits b_1', b_2', ... b_p' en résolvant un système de p équations normales à p inconnues (p = nombre de variables explicatives).

Prenons pour exemple le cas d'une régression à trois (3) variables explicatives (p=3), le système d'équations est le suivant :

$$r_{x1y} = b_1' + r_{x1x2}\, b_2' + r_{x1x3}\, b_3'$$
$$r_{x2y} = r_{x2x1}\, b_1' + b_2' + r_{x2x3}\, b_3' \quad \text{(III -3)}$$
$$r_{x3y} = r_{x3x1}\, b_1' + r_{x3x2}\, b_2' + b_3'$$

Ce système se résout par substitutions successives, tel que

$$b_1' = r_{x1y} - r_{x1x2}\, b_2' - r_{x1x3}\, b_3' \quad \text{(III-4)}$$

est placé dans les équations (2) et (3). On isole ensuite b_2' ou b_3' dans l'une des équations. Dès lors, on peut trouver une des valeurs b' et, en remontant la filière, on trouve les deux autres.

III.6.1.3 Coefficients de régression pour les variables d'origine

On trouve les coefficients de régression pour les variables originales b_1, b_2, ... b_p en multipliant chaque coefficient centré-réduit par l'écart-type de la variable dépendante (S_y), et en divisant le résultat par l'écart-type de la variable explicative considérée (S_{xi}).

Pour notre exemple à trois variables explicatives :

$$b_1 = b_1' S_y/S_{x1}$$
$$b_2 = b_2' S_y/S_{x2} \quad \text{(III-5)}$$
$$b_3 = b_3' S_y/S_{x3}$$

III.6.1.4 Ordonnée à l'origine

On trouve l'ordonnée à l'origine en posant la moyenne de la variable dépendante Y, et en lui soustrayant chaque coefficient de régression multiplié par la moyenne de la variable explicative correspondante:

$$b_0 = y - (b_1\bar{X}_1 + b_2\bar{X}_2 + ... + b_i\bar{X}_i ... + b_p\bar{X}_p) \quad \text{(III-6)}$$

Dans le cas de l'exemple précédent :

$$b_0 = y - b_1\bar{X}_1 - b_2\bar{X}_2 - b_3\bar{X}_3$$

III.7 Notation matricielle

Le système d'équations du modèle de régression peut être écrit sous la forme matricielle suivante

$$\begin{pmatrix} Y1 \\ Y2 \\ \vdots \\ Yn \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1p} \\ 1 & X_{21} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{np} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix} \dots\dots\dots\dots\dots\dots \text{(III-7)}$$

Ou plus simplement

$$Y = Xb + \varepsilon \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \text{(III-8)}$$

La somme des carrés des erreurs s'écrit

$$SCE = \varepsilon^T \varepsilon = (Y - Xb)^T (Y - Xb) \dots\dots\dots\dots\dots\dots\dots\dots \text{(III-9)}$$

Les b_i sont choisis de façon à minimiser les SCE. Le minimum est atteint lorsque toutes les dérivées partielles des SCE par rapport aux différents b_i s'annulent :

$$SCE = Y^TY - Y^TXb - b^TX^TY + b^TX^TXb \dots\dots\dots\dots\dots\dots\dots\dots \text{(III-10)}$$

$$\frac{\delta SCE}{\delta b} = (X^TX)b - X^TY = 0 \dots\dots\dots\dots\dots\dots\dots\dots \text{(III-11)}$$

D'où on tire finalement

$$b = (X^TX)^{-1} X^TY \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \text{(III-12)}$$

III.8 Qualité de la régression

Commençons d'abord par un tableau récapitulatif comportant les différentes notations qui seront utilisées dans ce qui suit :

Tableau III-1 : Récapitulatif somme des carrés

Nom	Sigle	Définition		DDL
SCT	Somme des carrés totale	$(Y-Y_{moy})^T (Y-Y_{moy})$	$\sum (Y_i - \bar{Y})^2$	n-1
SCR	Somme des carrés de la régression	$(Y_{sim}-Y_{moy})^T (Y_{sim}-Y_{moy})$	$\sum (\hat{Y}_i - \bar{Y})^2$	p
SCE	Somme des carrés des erreurs	$(Y-Y_{sim})^T (Y-Y_{sim})$	$\sum (Y_i - \hat{Y}_i)^2$	n-p-1

Y : Valeurs observées de la variable dépendante

$\bar{Y}$: Moyenne des valeurs observées de la variable dépendante

$\hat{Y}$: Valeurs simulées de la variable dépendante

n : Taille de l'échantillon.

p : Nombre de variables indépendantes.

Afin d'évaluer la qualité d'une régression multiple, les paramètres et tests suivants sont utilisés :

III.8.1 Coefficient de détermination multiple $\mathbf{R^2}$

Il permet de mesurer la variation expliquée par le modèle de régression à travers le rapport entre la dispersion expliquée par la régression (SCR) et la dispersion totale (SCT)

Le coefficient de détermination multiple est défini, lorsque le modèle de régression comporte une constante par la relation suivante :

$$R^2 = \frac{SCR}{SCT} = \frac{\sum_{i=1}^{n}(\hat{Y}i-\bar{Y})^2}{\sum_{i=1}^{n}(Yi-\bar{Y})^2} \quad \text{.....} \quad \text{(III-13)}$$

Cependant, l'expression suivante, prenant en compte les deux cas de présence et d'absence de constante est la plus utilisée :

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_{i=1}^{n}(Yi-\hat{Y}i)^2}{\sum_{i=1}^{n}(Yi-\bar{Y})^2} \quad \text{.....} \quad \text{(III-14)}$$

La variance totale étant la somme des variances expliquée et résiduelle

$$\text{SCT = SCE + SCR} \rightarrow \sum_{i=1}^{n}(Yi-\bar{Y})^2 = \sum_{i=1}^{n}(Yi-\hat{Y}i)^2 + \sum_{i=1}^{n}(\hat{Y}i-\bar{Y})^2 \quad \text{.....} \quad \text{(III-15)}$$

Les valeurs du R^2 varient dans l'intervalle [0, 1], mais contrairement à la régression simple, une valeur de R^2 proche de 1 n'est pas une indication suffisante à elle seule de la bonne qualité du modèle. Il est nécessaire d'effectuer un test sur la significativité de R afin de savoir s'il existe une relation entre Y et les X_i.

III.8.2 R^2 ajusté

Une des propriétés de la régression multiple est que l'ajout de chaque variable explicative au modèle permet d'"expliquer" plus de variation, et cela même si la nouvelle variable explicative est complètement aléatoire. Cela vient du fait que si l'on compare deux variables aléatoires, les fluctuations aléatoires de chacune d'entre elles produisent de très légères corrélations: Y et chacune des X_i ne sont pas strictement indépendantes (orthogonales) même s'il n'y a aucune relation réelle entre elles. Par conséquent, le R^2 comprend une composante déterministe, et une composante aléatoire d'autant plus élevée que le nombre de variables explicatives est élevé. Le R^2 est donc **biaisé**.

De plus, on ne peut pas comparer des modèles de complexité différente (nombre de variables indépendantes différent) sur la base du R^2, c'est pour ces raisons qu'un nouveau paramètre, le R^2 ajusté- noté aussi $\bar{R}^2$ – a été introduit par EZEKIEL (1930)

Son but est de « pénaliser » l'augmentation de la valeur du R^2 due à l'introduction de nouvelles variables dans le modèle. C'est un paramètre corrigé à l'aide des degrés de liberté. Il est donné par la relation suivante :

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} = 1 - \frac{SCE/n-p-1}{SCT/n-1} \quad \text{..... (III-16)}$$

Comme pour le R^2, les valeurs du $\bar{R}^2$ varient dans l'intervalle [0, 1]. Une valeur proche de 1 indique une bonne qualité de l'ajustement prenant en compte le nombre de variables explicatives. (*Rakotomalala*)

III.8.3 Test de signification du modèle de régression multiple

Un test statistique consiste à confronter les résultats d'une expérience à une hypothèse de départ (H_0). Pour réaliser un test, il faut connaître la distribution d'une statistique en supposant l'hypothèse de départ vérifiée.

Nous nous concentrerons sur le test le plus important en régression: est-ce que la pente de la régression est significativement différente de zéro ? (cette pente est égale à zéro lorsqu'il n'y a pas de relation entre les variables Y et X). Si les variables X expliquent vraiment Y alors la SCE sera faible car les erreurs seront faibles et SCR sera élevée.

III.8.3.1 Hypothèses du test

H_0 *:* $b_1 = b_2 = ... = b_p = 0$

La variable *Y* est linéairement indépendante des variables X_i. La régression est nulle, toutes les pentes du modèle sont égales à zéro.

$H_1 : \exists i \,/\, b_i \neq 0$

La variable *Y* est expliquée linéairement par au moins l'une des variables X_i. Une pente au moins est différente de zéro, la régression explique quelque chose.

La signification du modèle de régression multiple peut être testée par la variable auxiliaire *F* définie par :

$$F = \frac{R^2(n-p-1)}{p\,(R^2-1)} = \frac{SCR/p}{SCE/(n-p-1)} \quad \text{..... (III-17)}$$

L'hypothèse H_0 est rejetée au seuil α lorsque $F \geq F_{1-\alpha}(p,n-p-1)$

Ce test compare la variance expliquée avec celle des résidus. Si H_0 est vraie, ces deux valeurs devraient être à peu près semblables, et la statistique-test *F* suivra une distribution *F* de Fisher-Snedecor à *p* et (*n-p-1*) degrés de liberté.

Pratiquement, on calcule le rapport précédent et on le compare à la valeur lue dans la table. Si le rapport est supérieur à la valeur critique de la table c'est que la régression explique quelque chose et par conséquent H_0 est rejetée.

III.8.3.2 Conditions d'application du test

La régression multiple est soumise aux mêmes contraintes que la régression linéaire simple: distribution normale des résidus, équivariance, indépendance des observations et linéarité des relations entre la variable dépendante *Y* et chacune des variables explicatives X_i.

III.8.4 Validation du modèle de régression, étude des résidus

L'étude des résidus d'un modèle de régression vise plusieurs objectifs:

- Vérifier les postulats du modèle: normalité, homogénéité des variances des résidus (homoscédasticité) et indépendance des résidus.
- Détecter des données aberrantes qui s'écartent considérablement du modèle.
- Détecter des tendances particulières (ex. comportement quadratique des résidus) et des relations des résidus avec des variables externes qui permettraient d'affiner le modèle.

La **normalité** se vérifie essentiellement en construisant l'histogramme ou la fréquence cumulée des résidus. On peut vérifier l'ajustement à une normale visuellement ou effectuer des tests de normalité (ex. test d'ajustement du χ^2, test de Kolmogorov-Smirnov, etc...).

L'**indépendance des résidus** peut être testée en ordonnant les résidus en fonction d'un critère donné et en effectuant un test du genre: test des signes des résidus ou test de la corrélation entre résidus successifs dans la séquence ordonnée. Le test des signes (Draper et Smith, 1966; p.95) est un test non-paramétrique qui examine si l'arrangement des signes des résidus dans la séquence est aléatoire ou anormalement groupé ou encore anormalement fluctuant. Le test de corrélation consiste à calculer la corrélation entre les résidus et eux-mêmes décalés d'un pas dans la séquence. Si la corrélation est significative, alors il n'y a pas indépendance des résidus.

Le critère servant à ordonner la séquence peut être une variable interne (ex. la variable Y) ou une variable externe.

Cette prémisse peut aussi être vérifiée avec la statistique Durbin-Watson qui se situe entre 0 et 4, une valeur de 2 indiquant une absence de corrélation, moins de 2 une corrélation positive et plus de 2, une corrélation négative. Cette statistique est donnée par la formule suivante :

$$DW = \frac{\sum_{i=2}^{n}(\varepsilon i - \varepsilon i-1)^2}{\sum_{i=1}^{n} \varepsilon i^2} \quad \text{..... (III-18)}$$

L'**homogénéité des variances** des résidus se vérifie en ordonnant les résidus selon un critère comme ci-dessus et en vérifiant que les résidus montrent des variations de même amplitude pour toute la séquence ordonnée. Si ce n'est pas le cas, alors on peut tenter de corriger la situation à l'aide de transformations telles le logarithme ou la racine carrée qui ont habituellement pour effet de stabiliser la variance.

La **détection de données aberrantes** s'effectue en considérant les résidus qui s'écartent beaucoup de zéro. Les résidus situés à plus de trois écarts-types (l'écart-type des résidus est estimé par $\left(SCE \middle/ \sqrt{n-p-1}\right)$, sont suspects et doivent être examinés avec attention. Si des erreurs sont responsables de ces valeurs élevées, on doit les éliminer et reprendre la régression. Si aucune cause d'erreur ne peut les expliquer, alors il faut soit chercher à

affiner le modèle pour mieux expliquer ces données, soit chercher de nouvelles observations avec les mêmes valeurs de *X* que ces données pour en vérifier la validité.

La **détection de tendances particulières** dans les données se fait en reportant sur des diagrammes binaires les résidus en fonction de chacune des variables *X*. Des diagrammes binaires entre les résidus et des variables externes peuvent suggérer l'inclusion de nouvelles variables ou la transformation de variables existantes dans le modèle afin d'en améliorer la performance.

III.8.5 Multicolinéarité

Le terme multicolinéarité vise les phénomènes d'interdépendance (de corrélation) entre variables explicatives. On distinguera la multicolinéairté parfaite de la multicolinéarité partielle. Dans le premier cas, une variable explicative est une combinaison linéaire parfaite des autres variables explicatives. Dans le second cas, une variable explicative est fortement corrélée à une ou plusieurs variables explicatives (ou à l'une de leurs combinaisons **linéaires**)

III.8.5.1 Multicolinéarité parfaite

Dans ce cas, la matrice X est singulière et donc l'estimation des paramètres par MCO est impossible. Le problème de multicolinéarité parfaite est ainsi un **problème d'identification**. Les coefficients du modèle sont indéterminés et leur variance infinie.

III.8.5.2 Multicolinéarité partielle

Le cas le plus habituel rencontré avec les données est celui où les variables sont fortement mais pas parfaitement corrélée. Contrairement à la multicolinéarité parfaite, on ne va pas avoir un problème d'identification mais un **problème statistique** (**précision**). Les coefficients du modèle de régression peuvent être déterminés mais l'écart-type de leur estimation est important. (*Gujarati, 1995*)

III.8.5.3 Conséquences pratiques

En cas de multicolinéarité parfaite, la sanction est simple : l'algorithme des moindres carrés "plante".

En cas de multicolinéairté classique, on constate en général :

- que la variance de l'estimation des paramètres tend à être très forte
- que par conséquent, l'intervalle de confiance autour des paramètres s'élargit considérablement
- que l'estimation des paramètres est très sensible à la constitution de l'échantillon.

III.8.5.4 Solutions possibles

Afin de remédier au problème de multicolinéarité, il existe plusieurs solutions parmi lesquelles :

- Créer une nouvelle variable synthétique (combinant les variables inter-reliées) et l'utiliser à la place des autres
- Choisir une seule des variables très inter-reliées et s'en servir comme indicatrice des autres.

Remarque: si le seul but de la régression multiple est la prédiction (maximisation du R^2), la multicolinéarité ne dérange pas.

III.9 Types de régression multiple

Il existe plusieurs types de régression multiple, en fonction de l'introduction des variables explicatives dans le modèle. On distingue :

III.9.1 La régression hiérarchique

Cette méthode permet de déterminer l'ordre d'entrée des variables dans le modèle à l'aide de la création des blocs de variables qui seront entrés de manière hiérarchisée dans le modèle. Ceci permet d'observer plus en détail comment se comporte le modèle. Les résultats indiquent l'apport de chaque bloc en termes de pourcentage de variance expliquée (R^2). Pour les blocs constitués de plus d'une variable, il est possible de faire entrer celles-ci en un seul temps (entrée forcée) ou progressivement.

III.9.2 La régression avec entrée forcée

Toutes les variables explicatives sont introduites au même moment, et un test F évalue l'ensemble du modèle. L'ordre d'entrée des variables n'est pas influencé, le modèle évalue donc leur effet combiné.

III.9.3 La régression avec entrée progressive

Contrairement aux deux autres méthodes, la sélection des variables à inclure est basée sur un critère mathématique. Une fois les variables indépendantes choisies, leur inclusion dans le modèle dépendra de leur contribution mathématique à son amélioration. Il existe trois méthodes progressives.

III.9.3.1 Méthode rétrograde (backward selection)

Cette méthode consiste à construire un modèle de régression complet (intégrant toutes les variables explicatives), et à en retirer une par une les variables ayant la plus faible contribution au modèle (en commençant par celle qui explique le moins de variation). L'Inconvénient de cette méthode est qu'une fois qu'une variable a été retirée, elle ne peut plus être réintroduite dans le modèle, même si, à la suite du retrait d'autres variables, elle redevenait significative.

III.9.3.2 Méthode ascendante (forward selection)

Approche inverse de la précédente: elle sélectionne d'abord la variable explicative la plus corrélée à la variable dépendante. Ensuite, elle sélectionne, parmi celles qui restent, la variable explicative dont la corrélation partielle est la plus élevée (en gardant constantes la ou les variables déjà retenues). Et ainsi de suite tant qu'il reste des variables candidates dont le coefficient de corrélation partiel est significatif. L'Inconvénient est que lorsqu'une variable est entrée dans le modèle, aucune procédure ne contrôle si sa corrélation partielle reste significative après l'ajout d'une ou de plusieurs autres variables. Cette technique est en général plus conservatrice que la précédente, ayant tendance à sélectionner un modèle plus restreint (moins de variables explicatives) que la sélection rétrograde.

Toutefois, des simulations récentes (Blanchet *et al.* 2008 3) montrent que même la sélection progressive laisse souvent entrer au moins une variable non significative dans le modèle. C'est la raison pour laquelle un double critère d'arrêt à la sélection est appliqué :

1. Le niveau α habituel, et
2. Le $\bar{R}^2$ du modèle comprenant toutes les variables candidates.

Pour ce deuxième critère, on calcule tout d'abord le $\bar{R}^2$ global d'une régression multiple comprenant toutes les variables explicatives candidates. Ensuite, durant la procédure de sélection, on arrête la sélection lorsque le niveau α présélectionné **ou** le $\bar{R}^2$ global est atteint.

Cette procédure réduit fortement le nombre de variables explicatives introduites à tort dans le modèle.

III.9.3.3 La méthode pas-à-pas (stepwise regression)

Cette procédure, la plus complète, consiste à faire entrer les variables l'une après l'autre dans le modèle (selon leur corrélation partielle) par sélection progressive et, à chaque étape, vérifier si les corrélations partielles de l'ensemble des variables déjà introduites sont encore significatives (une variable qui ne le serait plus serait rejetée). Cette approche tente donc de

neutraliser les inconvénients des deux précédentes en les appliquant alternativement au modèle en construction.

III.10 Conclusion

Comme nous avons pu le constater à travers ce chapitre, la régression multiple est une technique de modélisation basée sur les moindres carrés ordinaires, ses objectifs sont variés et il existe plusieurs façons de procéder en fonction de l'introduction des différentes variables dans le modèle.

Afin d'évaluer la qualité du modèle, plusieurs critères tels que le R^2 et le R^2 *ajusté* sont utilisés. L'existence d'une relation entre la variable dépendante et les variables explicatives est quant à elle estimée par un test de Fisher.

Finalement, il faut faire attention aux différents problèmes que l'on peut rencontrer lors de la modélisation par régression multiple et ceci afin d'aboutir a un modèle robuste et de bonne qualité, permettant de faire des estimations ou des prédictions avec un niveau de précision appréciable.

CHAPITRE IV

MODELISATION DU COÛT DE REVIENT UNITAIRE DE L'EAU

IV.1 Introduction

Dans ce chapitre, nous allons mettre en pratique les méthodes détaillées auparavant, en les appliquant aux données concernant les barrages en exploitation, dans le but d'estimer le coût réel de l'eau et par conséquent évaluer le montant de la sujétion de service public, ce qui constitue le principal objet de ce travail.

IV.2 Les données de l'étude

Pour notre étude, nous nous basons sur le bilan comptable de l'exercice 2005-2006, cette année étant la seule où les différentes charges sont présentées de manière détaillée.

Le nombre de barrages en exploitation est de 64 barrages. Les données à disposition pour chaque barrage sont :

- Le total des charges directes
- Une part des charges indirectes
- La capacité de stockage
- Le volume mobilisé, c'est-à-dire celui vendu au prix unitaire de 1 DA/m3

  De plus, une seconde part de charges indirectes est représentée par le montant total résultant du fonctionnement du siège de l'ANBT ainsi que des différentes unités d'exploitation à travers le pays.

  Cette part de charges indirectes devra être imputée aux charges de chaque barrage en fonction d'une clé de répartition déterminée arbitrairement.

  La clé de répartition choisie est le volume mobilisé, autrement dit chaque barrage se verra imputer une part de ses charges indirectes au prorata de son volume mobilisé.

La clé de répartition pour chaque barrage aura donc pour formule :

$$\text{Clé de répartition} = \frac{Volume\ mobilisé\ du\ barrage}{Volume\ mobilisé\ total} = \frac{VM_i}{\sum_{i=1}^{n} VM_i}$$

Ceci étant, chaque barrage de notre tableau de données aura les caractéristiques suivantes :

- Charge directe notée « Cha.dir »
- Une première part des charges indirectes notée « Cha.ind.1 »
- Une clé de répartition permettant de lui imputer la seconde part des charges indirectes notée « clé.de.répart »

- Une seconde part de charges indirectes notée « Cha.ind.2 » qui représente la part des charges du siège et des unités d'exploitation supportée par le barrage en question.
- Le total des charges indirectes noté « Cha.ind » calculé comme suit :

$$Cha.ind = Cha.ind.1 + Cha.ind.2 \quad \text{(IV-1)}$$

- Capacité notée « Capacité »
- Volume mobilisé noté « Vol.mob »

  Nous pourrons donc, à partir de ces données, calculer le montant des charges totales pour chaque barrage en additionnant les charges directes et indirectes, cette charge totale notée « Cha.tot » est donc de :

$$Cha.tot = Cha.dir + Cha.ind \quad \text{(IV-2)}$$

  Ce calcul intermédiaire va nous permettre de calculer le coût de revient du mètre cube d'eau mobilisé. Il va sans dire qu'il varie d'un barrage à l'autre. Ce coût, noté « C.U. » est obtenu en divisant les charges totales par le volume mobilisé comme suit :

$$\text{C.U.} = \frac{Cha.tot}{Vol.mob} \quad \text{(IV-3)}$$

Le fichier de données final, à partir duquel l'étude sera menée, aura la forme de l'exemple suivant :

Tableau IV-1 : Exemple du tableau de données

Barrage	Capacité (m^3)	Cha.Dir (DA)	Cha.Ind.1 (DA)	Clé.de. répart	Cha.ind.2 (DA)	Cha.Ind (DA)	Cha.Tot (DA)	Vol.Mob (m^3)	C.U. (DA)
Ain Dalia	76080000,00	5044000,00	9708685,45	0,0141	5185269,35	14893954,80	19937954,80	11000000,00	1,81

Toutes les charges et les coûts sont exprimés en Dinars Algériens (DA)

Les volumes quant à eux, sont donnés en mètres cubes (m^3)

Remarque : La première part de charges indirectes était déjà imputée aux barrages dans le fichier de données original, sans mention de la clé de répartition utilisée à cet effet, pour la suite de ce travail, cette part à été gardée telle quelle et a été utilisée sans modification.

IV.2.1 Calcul du coût unitaire moyen

Le coût unitaire moyen servira de base dans l'évaluation de la sujétion de service public, cette dernière sera égale à ce coût unitaire moyen multiplié par le total des volumes mobilisés, à

laquelle on devra soustraire le montant obtenu grâce à la vente de ces volumes au prix administré de 1 DA/m^3.

Sujétion de service public = (Coût unitaire moyen x Total volumes mobilisés) – (1 x Total volumes mobilisés)

Avec :

Coût unitaire moyen = $\frac{1}{n}\sum_{i=1}^{n} C.U._i$

Total volumes mobilisés = $\sum_{i=1}^{n} Vol.mob_i$

n étant le nombre de barrages étudiés = 64 barrages.

De plus, le coût unitaire moyen, une fois comparé au coût unitaire standard moyen, permettra de valider et d'évaluer la précision des modèles qui seront établis.

Le coût unitaire moyen calculé pour l'ensemble des barrages est de

$C.U._{moy}$ = 3,19 DA

La sujétion de service public correspondante est alors de

Sujétion service public = 1 710 280 062,00 DA

IV.3 Analyse en composantes principales

Pour cette analyse, nous disposons d'un tableau de données rectangulaire dont les individus sont les 64 barrages en exploitation (*n*=64) et les variables sont les charges directes (Cha.dir), les charges indirectes (Cha.ind), les capacités (Capa), les volumes mobilisés (Vol.mob) et les coûts unitaires (C.U) (*p*=5).

IV.3.1 Objectifs

Le but recherché par l'application d'une ACP normée dans notre cas d'étude est double :

IV.3.1.1 Sur les variables

Nous verrons, à partir de la représentation des variables sur le cercle de corrélation, quelles sont celles qui sont fortement corrélées entre elles, et éliminer, si elle existe, toute redondance. Autrement dit, nous allons essayer de réduire le nombre de variables explicatives afin de simplifier l'application à posteriori d'une régression multiple.

IV.3.1.2 Sur les individus

L'ACP va nous permettre de voir, à travers la représentation des individus dans le nouvel espace déterminé par les axes factoriels, les barrages similaires et dissimilaires, ainsi que les barrages atypiques. Comme pour les variables, cette étape servira de préparation à l'application d'une régression multiple regroupant le plus grand nombre de barrages avec le meilleur ajustement possible.

IV.3.2 Pertinence de l'analyse

Avant de procéder à une Analyse en Composantes Principales, nous devons voir si une telle analyse s'adapte aux données disponibles pour cette étude.

Le tableau suivant résume les résultats concernant les tests de sphéricité de Bartlett ainsi que le KMO :

Tableau IV-2 : Indice KMO global et Sphéricité de Bartlett

Indice KMO	Signification de Bartlett
0,648	0,000

Comme nous pouvons le remarquer, le KMO avec une valeur de 0,648 appartient à l'intervalle [0,6 ; 0,7[ce qui, selon Kaiser (1974) représente une compressibilité moyenne de l'information.

De plus, le test de sphéricité de Bartlett avec une signification de 0,000 est très significatif, confirmant ainsi la possibilité de soumission des données à une ACP.

Nous pouvons aussi calculer les indices KMO par variable :

Tableau IV-3 : Indice KMO par variable

Variable	KMO_j
Cha.dir	0,995
Cha.ind	0,589
Capa.	0,581
Vol.mob	0,568
C.U.	0,766

Toutes les variables étudiées ont un KMO > 0,5, allant même jusqu'à une valeur de 0,995 pour ce qui est des charges directes, les conclusions obtenues grâce au KMO global et au test de sphéricité de Bartlett sont confortées.

IV.3.3 Extraction des composantes

Nous pouvons extraire autant de composantes que de variables c'est-à-dire 5.

Ci-dessous le tableau des poids factoriels, contenant les coefficients permettant d'exprimer chacune des variables en fonction des composantes extraites

Tableau IV-4 : Poids factoriels des variables

Variable	Composante				
	1	2	3	4	5
Cha.Dir	0,987	0,065	-0,144	0,034	0,000
Cha.Ind	0,999	0,009	0,052	-0,012	0,000
Capacité	0,985	0,064	-0,156	-0,028	0,000
Vol.Mob	0,966	-0,045	0,256	0,005	0,000
C.U.	-0,094	0,995	0,030	0,000	0,000

Nous constatons qu'excepté le C.U. qui est expliqué dans sa quasi-totalité par la seconde composante C2, toutes les variables sont grandement expliquées par la première composante C1.

IV.3.4 Choix du nombre de composantes

Il nous faut maintenant déterminer le nombre de composantes optimal, restituant le maximum d'information présente dans les variables initiales avec une perte minimale.

Pour ce faire, nous allons nous baser sur le critère de Kaiser, ainsi que la méthode graphique du coude de Cattell.

Le tableau suivant nous donne le pourcentage de variance expliquée par chaque composante ainsi que les pourcentages cumulés :

Tableau IV-5 : Variance et variance cumulée

Composante	Total	% de la variance	% cumulés
1	**3,883**	77,660	77,660
2	**1,001**	20,015	97,675
3	0,114	2,283	99,958
4	0,002	0,042	100,000
5	0,000	0,000	100,000

D’après Kaiser, on ne doit retenir que les composantes expliquant une variance supérieure à 1, dans notre cas il s’agit des composantes C1 et C2.

De plus, une représentation des composantes nous permet de constater le changement de pente brusque qui survient à partir de la 3éme composante, confirmant ainsi le nombre de deux composantes déterminé auparavant.

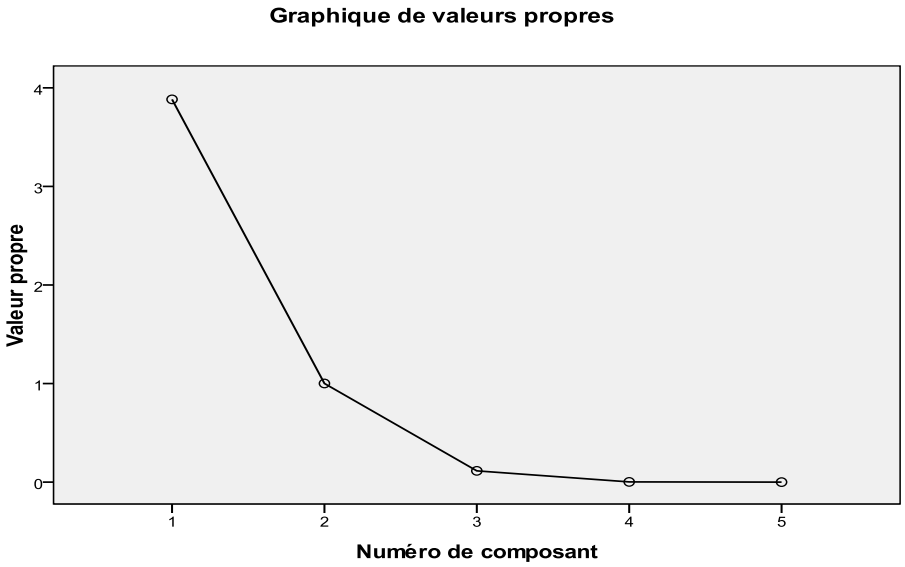

Figure IV-1 : Représentation des variances par composante

Nous constatons à partir du tableau des pourcentages de variance cumulée que les deux premières composantes extraites restituent à elles seules près de 98% de l’information totale, ce qui est considérable.

Nous pouvons donc nous arrêter à ces deux composantes, qui nous permettront également de représenter nos variables ainsi que les individus dans le plan factoriel correspondant.

IV.3.5 Représentation des variables

Les variables sont représentées sur le cercle de corrélation défini par les composantes C1 et C2 comme suit :

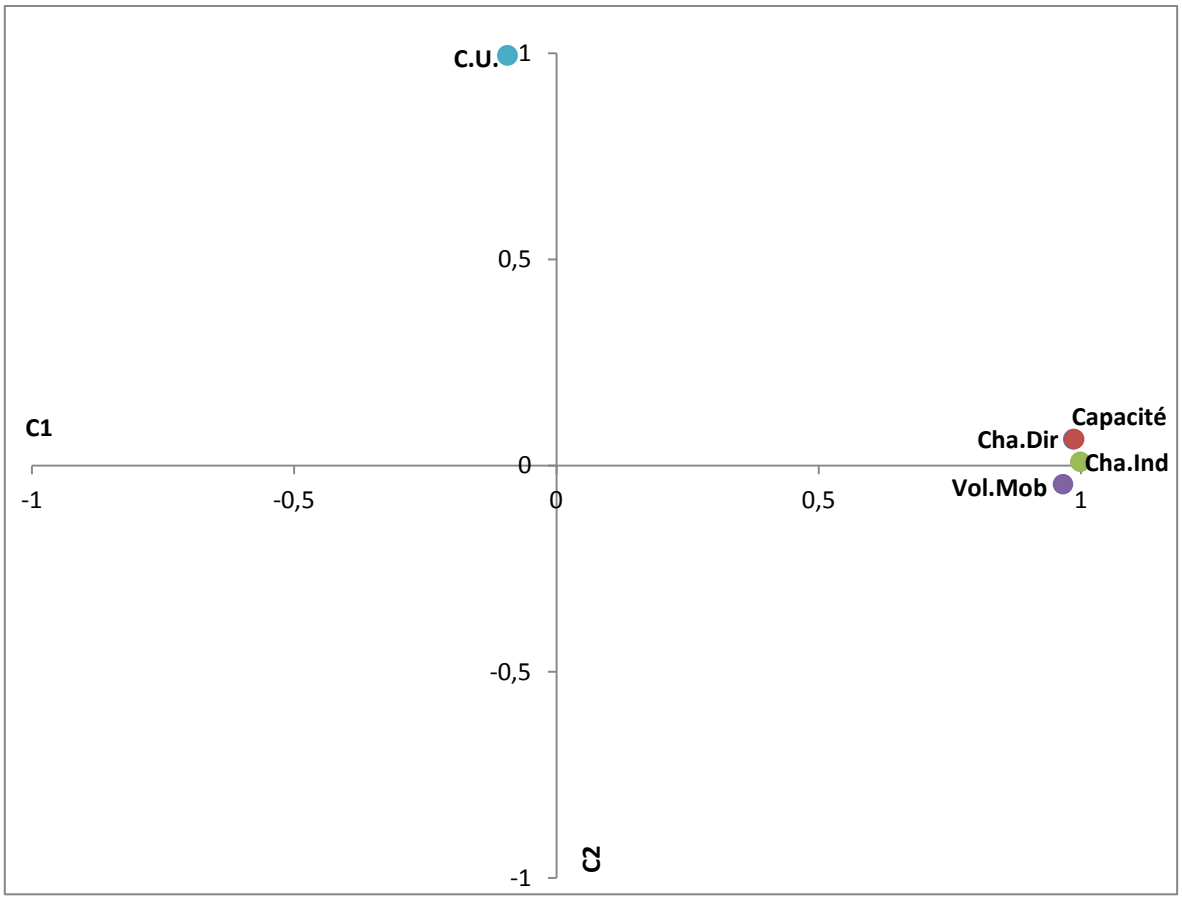

Figure IV-2 : représentation des variables

On remarque qu'il y a un groupement de quatre variables qui sont les charges directes, les charges indirectes, les capacités et les volumes mobilisés. Ces quatre variables contribuent fortement à la construction du premier axe factoriel, de plus, leur regroupement indique qu'elles expliquent la même chose, c'est-à-dire qu'elles renferment la même information.

Cette redondance est très utile pour la suite de notre étude car elle peut être exploitée dans la réduction du nombre de variables indépendantes dans une régression multiple, ou même dans la substitution d'une variable par une autre vu qu'elles expliquent la même chose.

IV.3.6 Représentation des individus

Les 64 barrages sont représentés dans le plan défini par les deux premiers axes factoriels correspondant aux composantes C1 et C2.

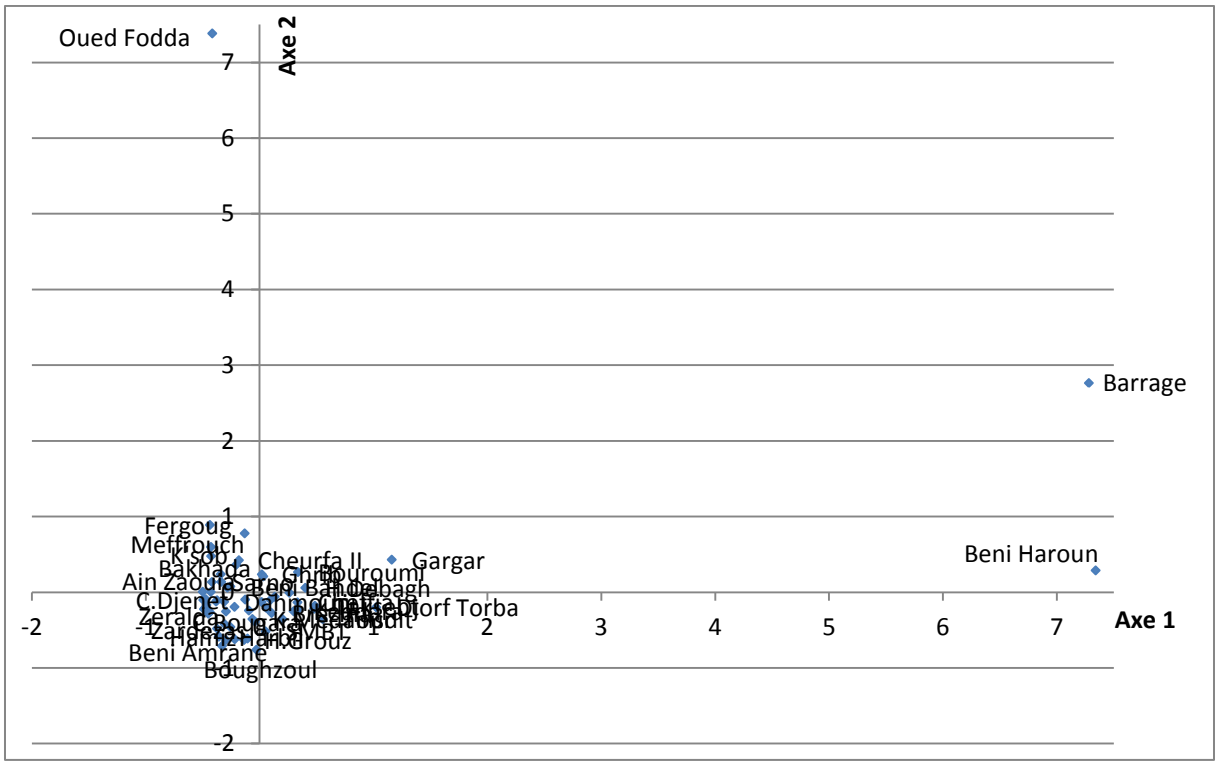

Figure IV-3 : Représentation des individus

La représentation des barrages dans le nouveau système d'axes factoriels considérés nous montre que le nuage de points se concentre autour de l'origine du repère, certains barrages tels que Beni Haroun, Oued Fodda, Gargar…etc s'éloignent quand à eux de cette origine.

Les coordonnées des barrages dans ce systèmes d'axes étant des valeurs adimensionnelles (car nous avons procédé à une ACP normée), nous pouvons en conclure que plus un barrage s'éloigne de l'origine (ou du centre de gravité du nuage qui sont quasiment confondus) plus ce barrage est différent de l'ensemble des autres barrages de par ses caractéristiques, que ça soit les charges, les volumes ou le coût, et par conséquent, sera susceptible de poser des problèmes lorsque l'on essayera d'établir un modèle de régression par la suite.

IV.3.7 Conclusions de l'analyse en composantes principales

1) La variable coût unitaire C.U. n'est corrélée avec aucune des autres variables, le modèle de régression comportera donc deux variables au moins.
2) Il existe une redondance évidente entre les variables : la capacité (Capa.), les charges directes (Cha.dir), les charges indirectes (Cha.ind) et dans une moindre mesure le volume mobilisé (Vol.mob) contiennent une même information.
3) On constate l'existence de barrages « atypiques » grâce à la représentation des individus dans le nouveau système d'axes factoriels.

IV.4 Régression multiple

Dans cette partie nous allons exploiter les conclusions de l'analyse en composantes principales afin d'établir le modèle le moins contraignant permettant de simuler le coût unitaire avec la plus grande facilité.

La variable dépendante est le coût unitaire, les variables indépendantes ou explicatives sont les charges directes, les charges indirectes, les capacités et les volumes mobilisés.

Bien que la loi permettant de calculer directement le coût de revient unitaire soit connue (C.U. = $\frac{Cha.Dir + Cha.ind}{Vol.mob}$), le but de notre approche est de trouver le meilleur moyen de simuler le coût unitaire en fonction d'un nombre minimal de variables et à partir des variables explicatives les plus simples à déterminer.

L'idée de base est, dans un premier temps, d'écarter si possible la variable Cha.ind car elle est, de toutes les variables, celle qui est la plus difficile à déterminer comme expliqué dans le premier chapitre de cette étude.

Le fait de devoir passer par des calculs intermédiaires (choix d'une clé de répartition, imputation) est le principal moteur de ce choix.

Dans ce sens, la redondance observée grâce à l'analyse en composantes principales est encourageante et laisse penser qu'une simulation du coût unitaire écartant les charges indirectes est envisageable.

Par la suite, nous essayerons de réduire le nombre de variables explicatives de notre modèle de 3 à 2 variables.

Le type de régression choisi est la régression hiérarchique car elle permet de sélectionner les variables d'entrée.

Une comparaison entre le coût de revient unitaire réel et celui simulé est, dans notre cas, le meilleur indicateur de la qualité des modèles établis. Nous nous baserons sur cette comparaison, en plus des critères dévaluation, dans la validation de nos modèles.

IV.4.1 Modèles de la forme $Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots b_p X_p + \varepsilon$

Avant d'entamer la réduction du nombre de variables, nous devons voir si un modèle de cette forme donne des résultats satisfaisants.

Les variables explicatives entrant en jeu dans le calcul du coût unitaire sont les charges directes, les charges indirectes et le volume mobilisé. Ces trois variables serviront de base dans l'établissement de notre premier modèle.

Le modèle de régression établi est le suivant :

$$C.U. = 2{,}649 + 4{,}962 * 10^{-7} (Cha.dir) - 1{,}436 * 10^{-7} (Cha.ind) + 7{,}084 * 10^{-9} (Vol.mob)$$

IV.4.1.1 Qualité de la régression

Les paramètres permettant d'évaluer le modèle sont présentés dans le tableau suivant :

Tableau IV-6 : Critères d'évaluation

R^2	R^2 ajusté	F
0,06	0,013	1,286

Il est clair qu'avec une valeur de R^2 proche de zéro, le modèle sous sa forme actuelle est rejeté sans procéder à une analyse plus approfondie.

IV.4.1.2 Elimination des valeurs extrêmes

La piètre qualité du modèle indique la présence de barrages dont le coût n'a pu être simulé qu'avec une grande erreur, c'est-à-dire s'éloignant trop de la droite de régression.

Afin de détecter et d'écarter ces barrages, nous allons procéder à partir de la représentation du nuage des individus dans le plan factoriel défini par les deux composantes issues de l'ACP.

Chaque barrage étant défini par ses coordonnées, adimensionnelles, nous écartons le barrage le plus éloigné du centre de gravité du nuage en terme de distance, et ainsi de suite jusqu'à stabilisation du R^2 à une valeur maximale (maximisation du R^2).

Le centre de gravité du nuage a pour coordonnées les moyennes arithmétiques des coordonnées factorielles :

$$X_G = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ et } Y_G = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Où X_i : Coordonnée du barrage i sur le premier axe factoriel

Y_i : Coordonnée du barrage i sur le second axe factoriel

n : Nombre de barrages = 64

Par conséquent, la distance entre un individu et le centre de gravité du nuage de points est définie comme :

$$D_i = \sqrt{(X_i - X_G)^2 + (Y_i - Y_G)^2}$$

Les distances ainsi calculées permettent d'écarter un groupe de 12 barrages qui sont :

Tableau IV-7 : Barrages exclus de l'analyse

Barrage	D
Oued Fodda	7,396
Beni Haroun	7,346
Gargar	1,238
Djorf Torba	1,042
Fergoug	0,991
Boughzoul	0,759
Meffrouch	0,736
Harbil	0,687
Hamiz	0,684
Tilisdit	0,658
K'sob	0,640
Zardezas	0,608

Les deux groupes de barrages résultant de cette démarche, un premier de 52 barrages, le second de 12 barrages, seront traités séparément comme suit :

IV.4.1.2.1 Groupe I :

Pour ce premier groupe, comportant un total de 52 barrages, les résultats de la régression donnent le modèle suivant, appelé modèle I :

$$C.U. = 3{,}005 + 1{,}351 * 10^{-7}\ (Cha.dir) + 1{,}016 * 10^{-7}\ (Cha.ind) - 2{,}947 * 10^{-7}\ (Vol.mob)$$

Les caractéristiques du modèle sont :

Tableau IV-8 : Critères d'évaluation modèle I

R^2	R^2 ajusté	F	D-W
0,743	0,727	46,245	1,627

Avec un $\bar{R}^2$ de 0,73 le modèle est satisfaisant, la valeur de la statistique de Fisher qui est appréciable va dans ce sens.

L'étude des résidus montre qu'ils présentent une légère corrélation qui est toutefois acceptable, comme l'atteste la valeur de la statistique de Durbin-Watson appartenant à l'intervalle [1, 3].

La normalité des résidus est quant à elle vérifiée par l'histogramme que voici :

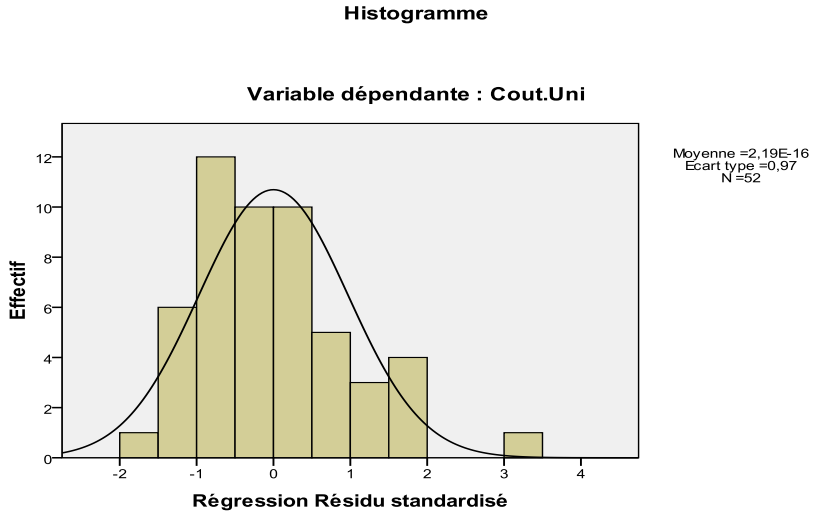

Figure IV-4 : Histogramme des résidus

IV.4.1.2.2 Groupe II :

Sur ce second groupe constitué de 12 barrages, la même démarche ne mène pas à une loi générale, quelles que soient les variables explicatives sélectionnées la valeur du R^2 ne dépasse pas 0,080.

Il faut encore exclure le barrage de Oued Fodda si l'on veut aboutir à un modèle satisfaisant.

La loi de régression obtenue pour le groupe de 11 barrages restants, appelée modèle II est la suivante :

$$C.U. = 0{,}949 + 4{,}262 * 10^{-6}\,(Cha.dir) - 2{,}137 * 10^{-6}\,(Cha.ind) + 9{,}761 * 10^{-7}\,(Vol.mob)$$

Elle présente les caractéristiques suivantes :

Tableau IV-9 : Critères d'évaluation modèle II

R^2	R^2 ajusté	F	D-W
0,826	0,751	11,079	1,888

De même que pour le premier groupe, ce modèle présente des qualités satisfaisantes sans être toutefois exceptionnelles.

Le barrage de Oued Fodda, n'appartenant à aucun groupe, devra être traité de façon individuelle.

IV.4.1.2.3 Réduction du nombre de variables

Bien que les deux modèles obtenus présentent de bonnes caractéristiques, leur intérêt sous leur forme actuelle est nul, vu que les variables explicatives qu'ils comportent sont les charges directes, les charges indirectes et le volume mobilisé.

Or, si nous disposons des informations détaillées concernant ces trois variables, le coût de revient pourra être calculé directement sans passer par un modèle de régression. C'est pour cette raison qu'il est impératif de réduire le nombre de variables en éliminant les charges indirectes, ou du moins, remplacer les charges indirectes par la variable capacité. Les résultats pour les deux groupes sont les suivants :

IV.4.1.2.3.1 Groupe I

Il est possible de se passer des charges indirectes quasiment sans répercussion sur la qualité du modèle, la nouvelle droite de régression appelée modèle I' est :

$$C.U. = 2{,}981 + 3{,}355 * 10^{-7} (Cha.dir) - 2{,}467 * 10^{-7} (Vol.mob)$$

Tableau IV-10 : Critères d'évaluation modèle I'

R^2	R^2 ajusté	F	D-W
0,732	0,721	66,855	1,601

De plus, la redondance tirée de l'analyse en composantes principales nous permet de remplacer la variable charges directes par la variable capacité, ce qui nous mène pour ce groupe au modèle final I'' :

$$C.U. = 3{,}026 + 2{,}121 * 10^{-8} (Capa) - 2{,}430 * 10^{-7} (Vol.mob)$$

Tableau IV-11 : Critères d'évaluation modèle I''

R^2	R^2 ajusté	F	D-W
0,739	0,728	69,287	1,646

Les valeurs du R^2, du $\bar{R}^2$ et du F sont satisfaisantes, l'analyse des résidus montre qu'ils sont indépendants (DW =1,646) et suivent une loi normale comme on peut le constater sur l'histogramme :

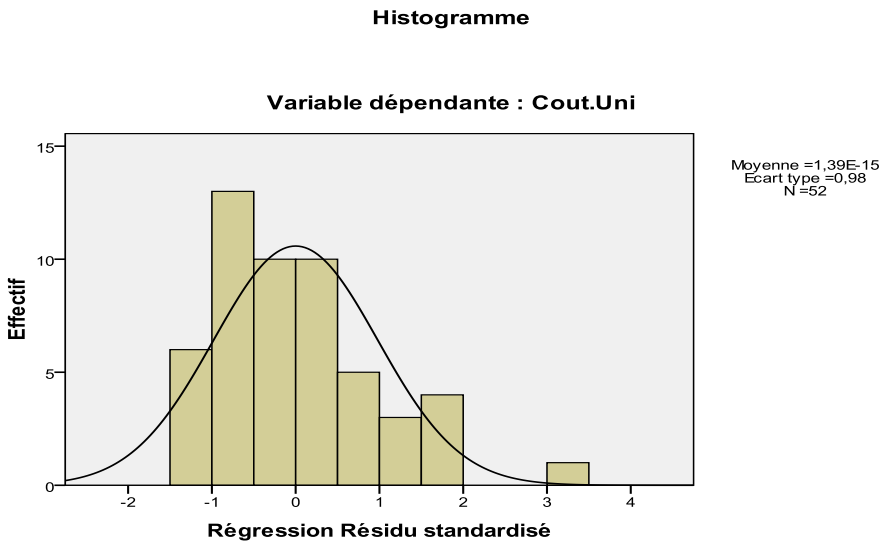

Figure IV-5 : Histogramme des résidus modèle I''

IV.4.1.2.3.2 Groupe II

Il est impossible pour ce groupe de modifier le modèle obtenu précédemment que ça soit en nombre de variables explicatives ou en nature des variables entrant en jeu dans la prédiction.

Pour exemple, l'élimination de la variable charges indirectes mène à la chute du $\bar{R}^2$ de la valeur 0,751 à une valeur de 0,150 indiquant une piètre qualité du modèle.

Cette contrainte, ne pouvant être contournée, nous oblige à chercher une nouvelle forme de régression multiple, car, comme le modèle du second groupe implique la connaissance des variables Cha.dir, Cha.ind et Vol.mob, il devient par conséquent obsolète, le coût de revient pouvant être calculé sans passage par un modèle.

De plus, le barrage de Oued Fodda, dont le coût de revient ne peut être simulé par les modèles ci-dessus doit être soumis à une étude séparée.

IV.4.1.2.4 Conclusion

- Il n'existe pas de modèle de régression de la forme $Y = b_0 + b_1 X_1 + b_2 X_2 \ \ldots b_p X_p + \varepsilon$ regroupant l'ensemble des barrages étudiés.
- Afin d'obtenir des résultats satisfaisants il faut séparer les barrages en deux groupes.

- Le modèle de régression du second groupe (de 11 barrages) doit inclure les trois variables intervenant dans le calcul direct du coût de revient, il en perd toute utilité pratique.
- Le barrage de Oued Fodda est exclu des modèles établis pour chaque groupe.
- Au vu des différents obstacles rencontrés, la piste menant aux modèles de la forme $Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots b_p X_p + \varepsilon$ doit être abandonnée. Il faut trouver une loi de régression différente.

IV.4.2 Modèles de la forme $Y = b_0 X_1^{b1} X_2^{b2} \ldots X_p^{bp} + \varepsilon$

Cette forme de régression multiple a notamment été utilisée par l'USGS (United States Geological Survey) dans l'estimation de fréquences des inondations.

Il est possible de linéariser cette équation par l'introduction d'un logarithme comme suit :

Prenons l'exemple d'une régression à deux variables explicatives X_1 et X_2.

$Y = b_0 X_1^{b1} X_2^{b2} + \varepsilon$

$Ln\ (Y) = Ln\ (b_0 X_1^{b1} X_2^{b2}) + \varepsilon$

$Ln\ (Y) = Ln\ (b_0) + Ln\ (X_1^{b1}) + Ln\ (X_2^{b2}) + \varepsilon$

$Ln\ (Y) = Ln\ (b_0) + b_1 Ln\ (X_1) + b_2 Ln\ (X_2) + \varepsilon$

C'est une équation de la forme

$Y' = b_0' + b_1 X_1' + b_2 X_2' + \varepsilon$

Cela revient à faire une régression multiple classique sur les nouvelles variables qui sont

$Y' = Ln\ (Y)$

$X_1' = Ln\ (X_1)$

$X_2' = Ln\ (X_2)$

A noter que les termes b_0 et ε sont des termes généraux représentant respectivement la constante et l'erreur de la régression.

D'une manière générale, une régression multiple de la forme

$$Y = b_0 X_1^{b1} X_2^{b2} \ldots X_p^{bp} + \varepsilon$$

n'est autre qu'une régression linéaire multiple de la forme

$$Y' = b_0 + b_1 X_1' + b_2 X_2' + \ldots b_p X_p' + \varepsilon$$

avec $Y' = Ln\ (Y)$, $X_i' = Ln\ (X_i)$ $i = 1, \ldots, p$

IV.4.2.1 Modèle A : C.U. = b_0 (Cha.dir) ^{b1} (Cha.ind) ^{b2} (Vol.mob) ^{b3}

La première réflexion est de vérifier, comme pour le modèle de régression linéaire classique, si le coût de revient peut être simulé par les trois variables permettant de le calculer naturellement, à savoir les charges directes, les charges indirectes et le volume mobilisé, et ce, pour l'ensemble des barrages.

Le calcul des coefficients de régression nous donne le résultat suivant :

$$C.U. = 2{,}617 * Cha.Dir^{\,0{,}435} * Cha.Ind^{\,0{,}465} * Vol.Mob^{\,-0{,}909}$$

IV.4.2.1.1 Evaluation du modèle

Les caractéristiques de l'ajustement sont résumées dans le tableau suivant

Tableau IV-12 : Critères d'évaluation modèle A

R^2	R^2 ajusté	F	D-W
0,988	0,988	1680,789	1,946

Ce modèle, avec une $\bar{R}^2$ de près de 0,99 présente d'excellents résultats, confirmés par une valeur de la statistique F très élevée.

Les résidus sont indépendants (DW ≈ 2) et suivent une loi normale comme nous pouvons le voir sur l'histogramme des résidus

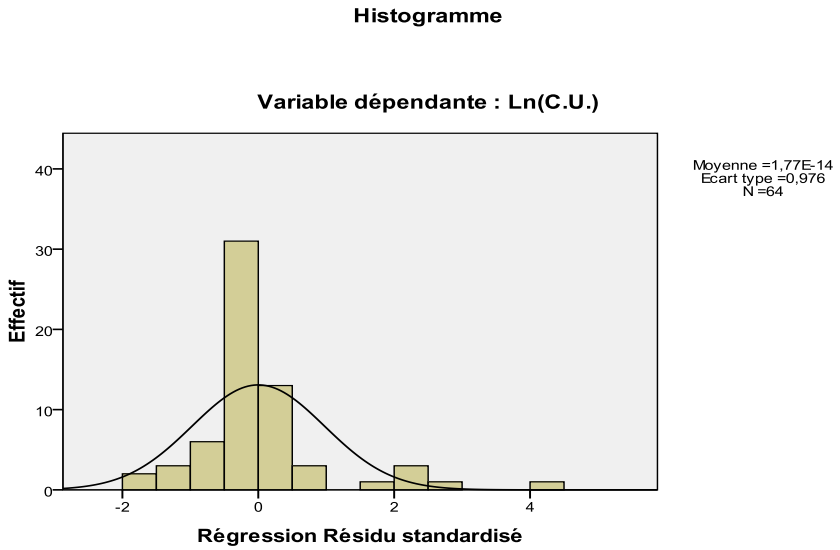

Figure IV-6 : Histogramme des résidus modèle A

Les résultats de la simulation sont présentés sous forme d'histogramme :

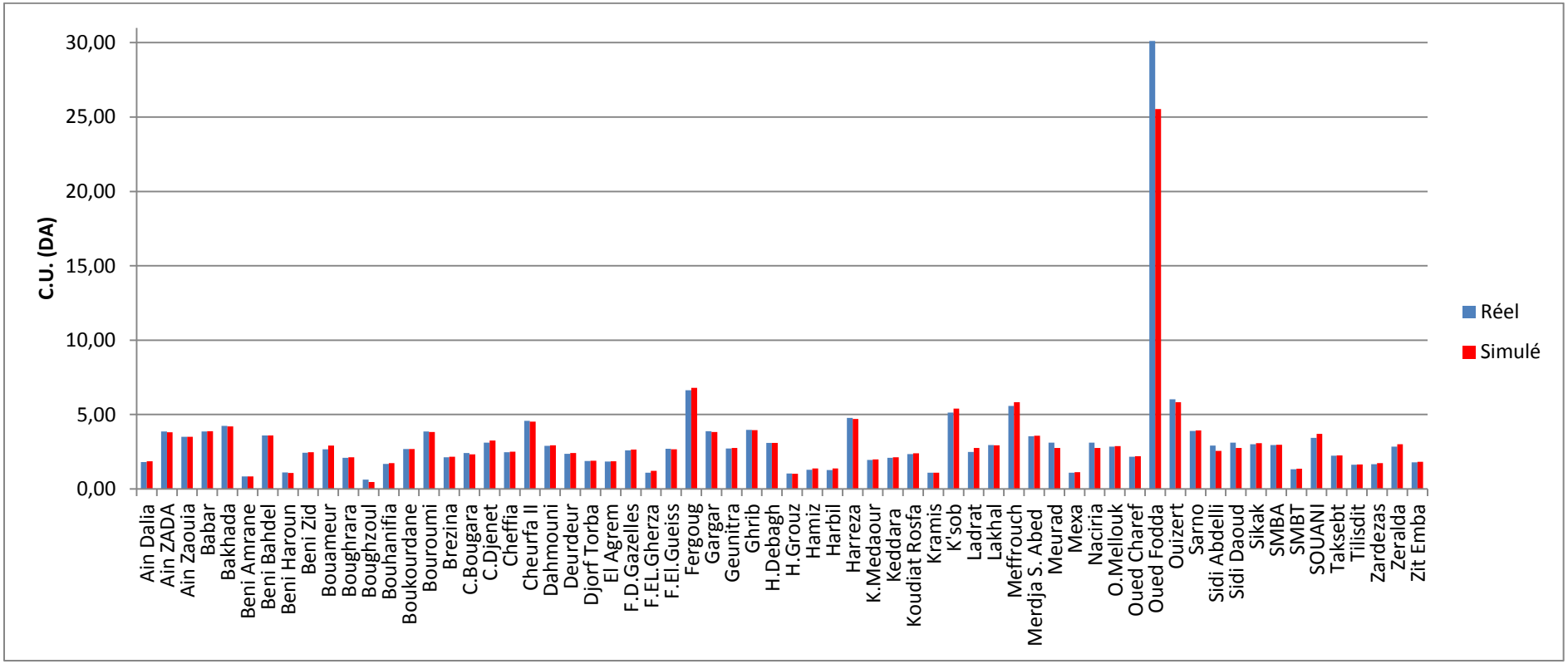

Figure IV-7 : Résultats simulation modèle A

Tableau IV-13 : Coût de revient unitaire moyen modèle A

$C.U._{Moy}$ réel (DA)	$C.U._{Moy}$ Simulé (DA)
3,19	3,13

Nous remarquons que les coûts unitaires simulés correspondent parfaitement aux coûts réels, exception faite du barrage de Oued Fodda où l'on constate une sous estimation du résultat simulé, c'est cette sous estimation qui est la principale cause de l'écart entre le coût unitaire moyen et celui simulé (bien que cet écart soit très petit).

IV.4.2.2 Modèle B : C.U. = b_0 (Cha.dir) ^{b1} (Capa) ^{b2} (Vol.mob) ^{b3}

Comme nous avons pu l'observer suite à l'analyse en composantes principales, la redondance dans les variables va nous permettre de substituer, dans le modèle précédent, les capacités aux charges indirectes.

L'idée est d'écarter les charges indirectes, difficiles à « évaluer » afin d'obtenir un modèle d'estimation plus simple à utiliser, ne comportant que des variables plus ou moins faciles à déterminer.

Le modèle obtenu est le suivant :

$$C.U. = 1{,}29 * Cha.Dir^{\,0{,}481} * Capa^{\,0{,}267} * Vol.Mob^{\,-0{,}733}$$

IV.4.2.2.1 Evaluation du modèle

Tableau IV-14 : Critères d'évaluation modèle B

R^2	R^2 ajusté	F	D-W
0,956	0,954	435,047	1,980

On remarque que même en remplaçant les charges indirectes par les capacités, le modèle donne d'excellents résultats comme en attestent les valeurs élevées du $\bar{R}^2$ et de F.

Les résidus quant à eux sont indépendants (très faible corrélation négative) et suivent une loi normale.

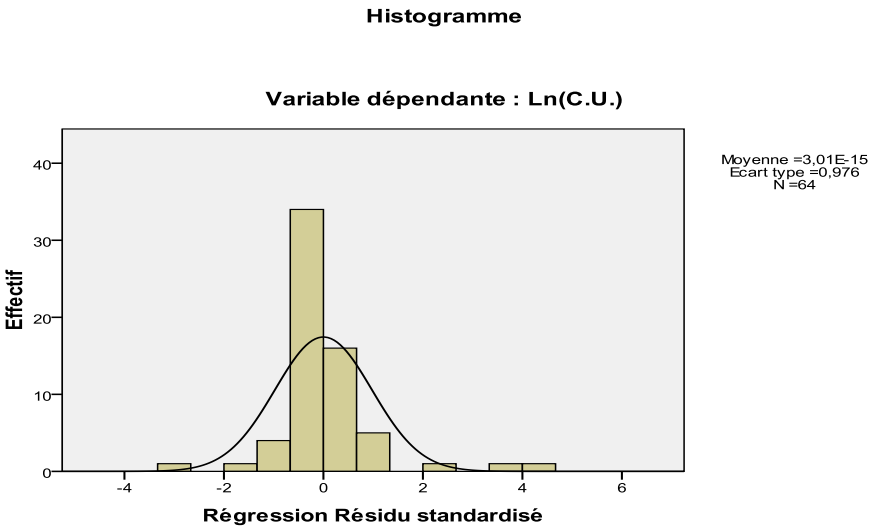

Figure IV-8 : Histogramme des résidus modèle B

Les résultats de la simulation sont représentés comme suit

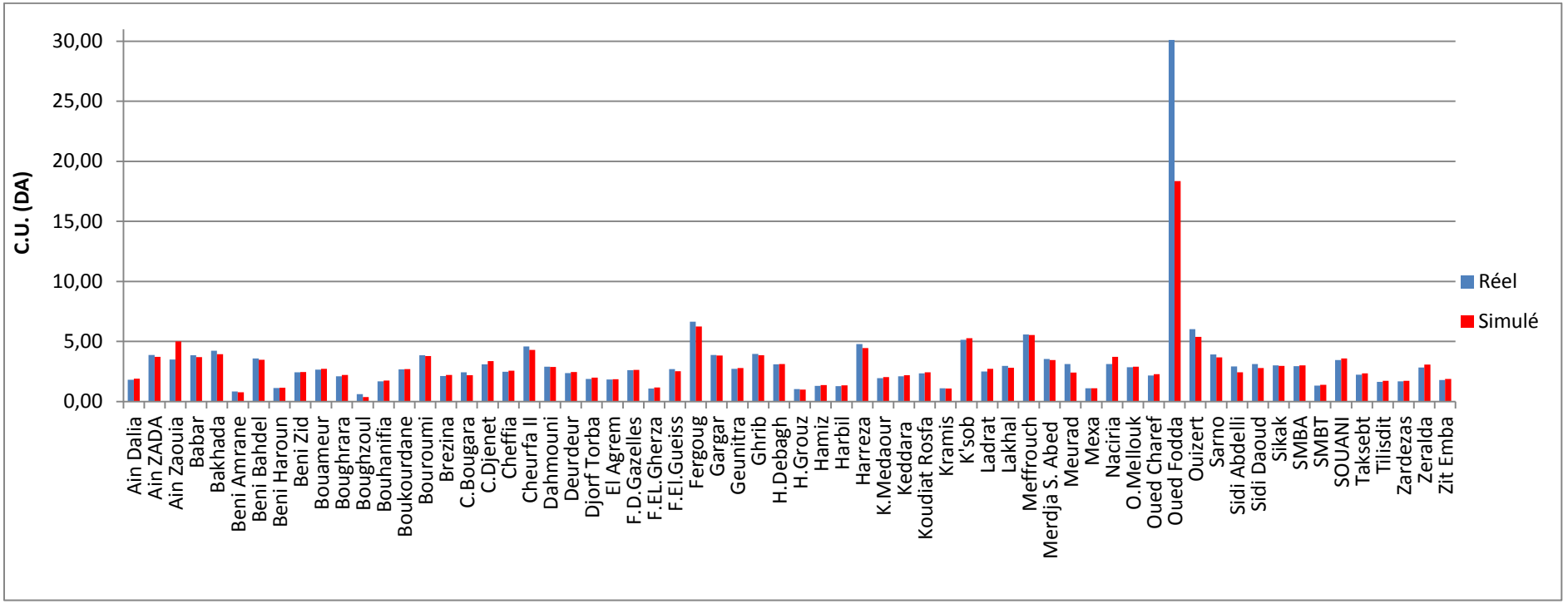

Figure IV-9 : Résultats simulation modèle B

Tableau IV-15 : Coût de revient unitaire moyen modèle B

$C.U._{Moy}$ réel (DA)	$C.U._{Moy}$ Simulé (DA)
3,19	3,00

Comme pour le modèle A, mis à part le barrage de Oued Fodda, les résultats de la simulation sont excellents, néanmoins, la grande sous estimation du coût unitaire de Oued Fodda se ressent sur la moyenne simulée qui voit sa valeur diminuer.

IV.4.2.3 Modèle B' :

L'écart entre la moyenne du coût de revient simulée et calculée est essentiellement dû à la grande sous estimation du coût de revient concernant le barrage de Oued Fodda (près de 12 DA de différence).

C'est pour cette raison que nous allons écarter ce barrage de l'étude, son coût de revient unitaire sera calculé de façon classique en utilisant la loi (C.U. = $\frac{Cha.Dir + Cha.ind}{Vol.mob}$) et c'est cette valeur qui sera prise en compte dans l'estimation du coût de revient unitaire moyen pour l'ensemble des barrages.

Le modèle obtenu dans ce cas est le suivant :

$$CU = 1{,}397 * Cha.Dir^{0{,}457} * Capa^{0{,}244} * Vol.Mob^{-0{,}68}$$

IV.4.2.3.1 Evaluation du modèle

Les critères d'évaluation du modèle sont

Tableau IV-16 : Critères d'évaluation modèle B'

R^2	R^2 ajusté	F	D-W
0,958	0,956	446,331	2,066

L'exclusion du barrage de Oued Fodda conduit à une légère amélioration des critères d'évaluation du modèle qui étaient déjà excellents. Les résidus sont indépendants et suivent une loi normale.

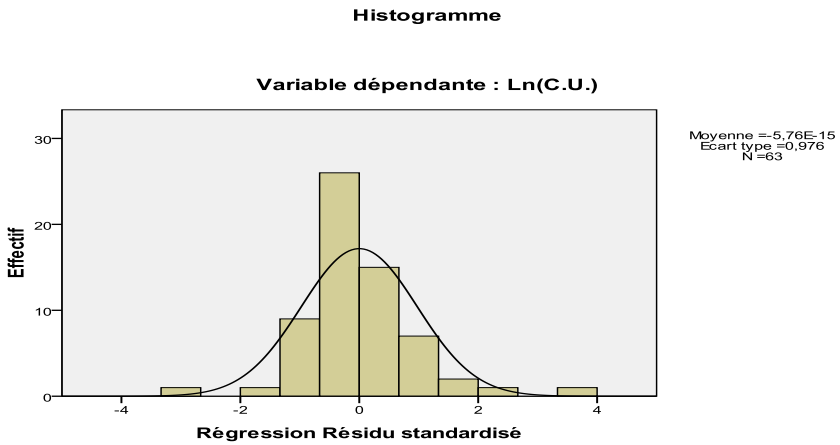

Figure IV-10 : Histogramme des résidus modèle B'

Cette tendance est confirmée par les résultats de la simulation

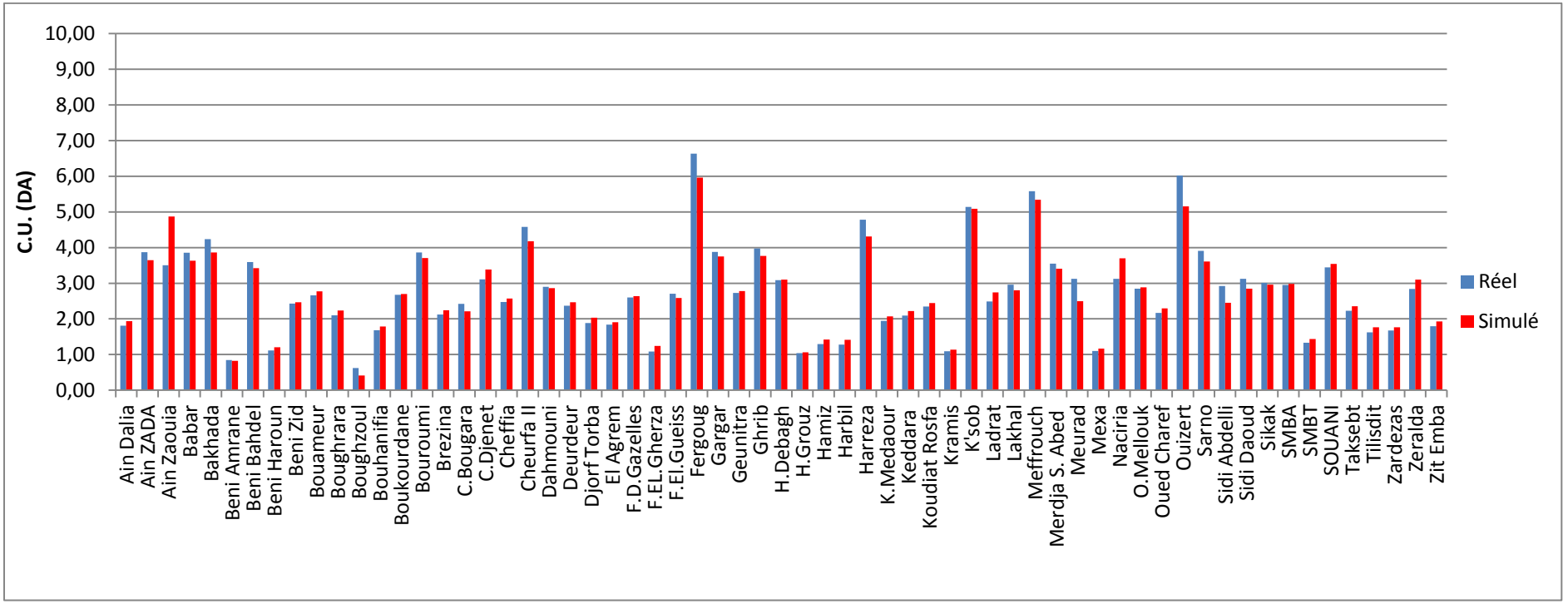

Figure IV-11 : Résultats simulation modèle B'

Tableau IV-17 : Coût de revient unitaire moyen modèle B'

	$C.U._{Moy}$ réel (DA)	$C.U._{Moy}$ Simulé (DA)
Oued Fodda exclu	2,76	2,75
Oued Fodda réintroduit	3,19	3,17

L'exclusion du barrage de Oued Fodda améliore grandement les résultats de la simulation, cette amélioration se voit principalement sur la moyenne simulée qui est quasiment égale à la moyenne réelle confirmant ainsi la grande influence du barrage de Oued Fodda sur les résultats finaux.

IV.4.2.4 Modèle C : C.U. = b_0 (Cha.dir) ^{b1} (Vol.mob) ^{b2}

Cette étape, toujours dans le but d'atteindre le modèle le plus simple, se concrétise par la réduction du nombre de variables à partir du modèle précédent.

Les conclusions de l'analyse en composantes principales affirment que certaines variables renferment la même information, est-il possible dans ce cas, d'éliminer l'une de ces variables sans altérer la qualité du modèle ?

Le modèle résultant de cette réflexion ne comporte que les charges directes et les volumes mobilisés, son équation est la suivante :

$$C.U. = 0{,}523 * Cha.Dir^{0{,}762} * Vol.Mob^{-0{,}645}$$

IV.4.2.4.1 Evaluation du modèle

Tableau IV-18 : Critères d'évaluation modèle C

R^2	R^2 ajusté	F	D-W
0,878	0,874	219,617	2,145

Bien que la valeur du $\bar{R}^2$ baisse par rapport aux deux modèles précédents, elle reste néanmoins appréciable tout comme la valeur de F.

De même, l'indépendance et la normalité des résidus sont toujours vérifiées.

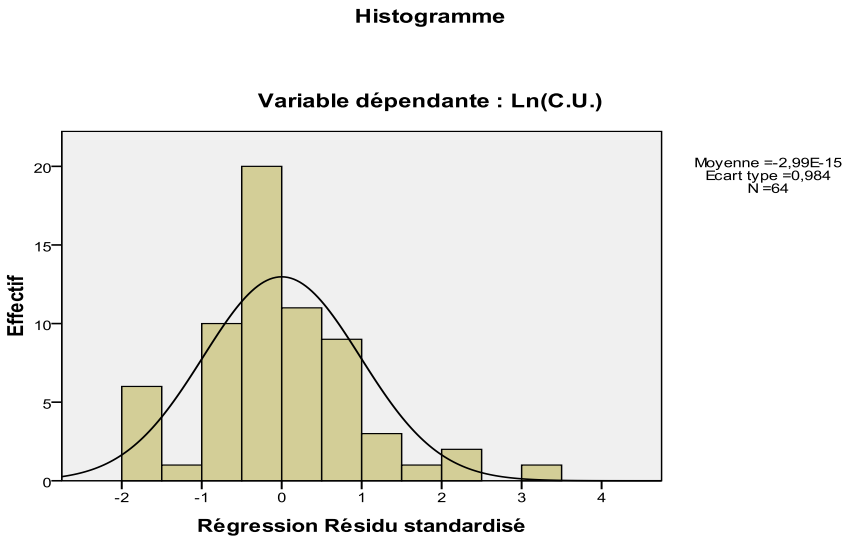

Figure IV-12 : Histogramme des résidus modèle C

La confrontation entre les coûts simulés et réels est visible sur l'histogramme suivant :

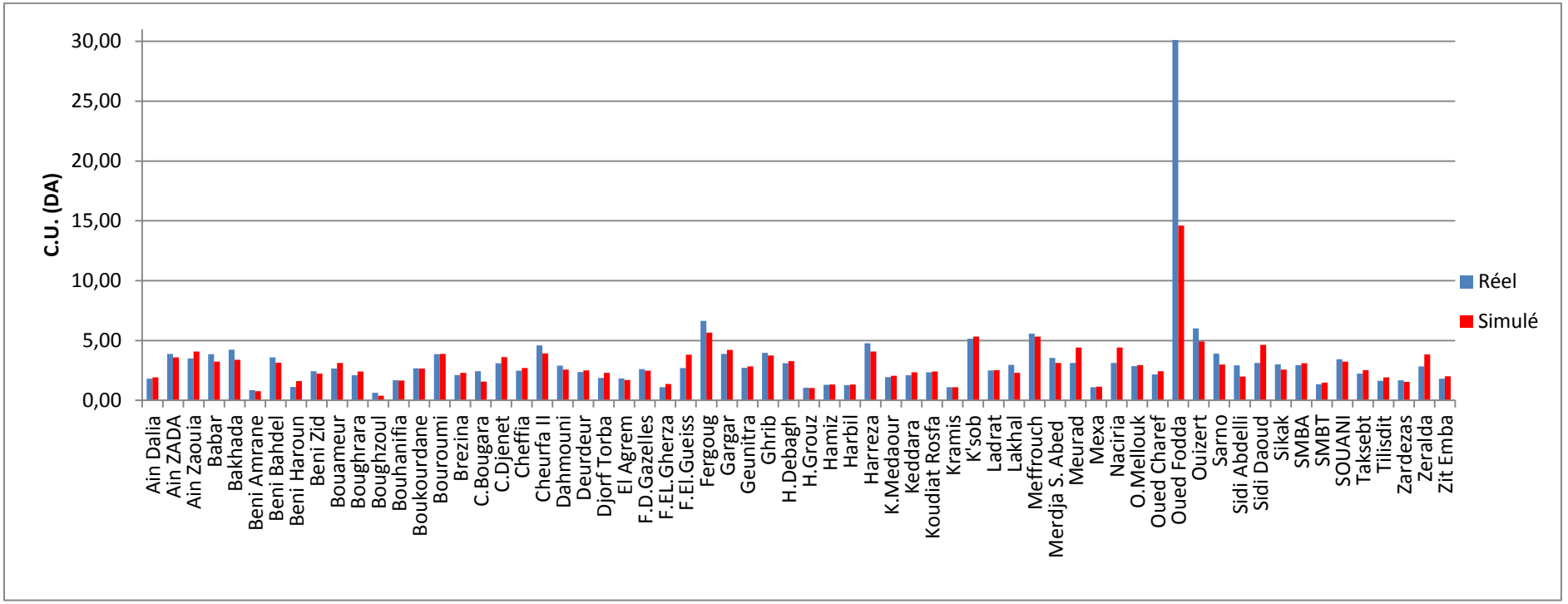

Figure IV-13 : Résultats simulation modèle C

Tableau IV-19 : Coût de revient unitaire moyen modèle C

$C.U._{Moy}$ réel (DA)	$C.U._{Moy}$ Simulé (DA)
3,19	2,96

L'énorme écart entre les valeurs simulée et calculée de la moyenne des coûts de revient unitaires –qui même s'il n'est que de 0,23 DA, représentera une somme élevée une fois multiplié par le nombre total de mètres cubes mobilisés- est expliqué par la sous estimation du coût de revient unitaire du barrage de Oued Fodda à elle seule. Le modèle reproduit assez bien les valeurs du coût unitaire pour tous les autres barrages.

IV.4.2.5 Modèle C'

Ce modèle, variante du modèle précédent, exclut le barrage de Oued Fodda de l'analyse car comme nous avons pu le voir, il est le principal responsable de l'écart existant entre le coût moyen réel et celui simulé.

Il est donné par la formule :

$$C.U. = 0{,}657 * Cha.Dir^{0{,}692} * Vol.Mob^{-0{,}592}$$

IV.4.2.5.1 Evaluation du modèle

Tableau IV-20 : Critères d'évaluation modèle C'

R^2	R^2 ajusté	F	D-W
0,870	0,866	201,161	2,188

Le modèle C' présente pratiquement les mêmes caractéristiques que le modèle C, que ce soit en terme de qualité de l'ajustement ou en terme d'indépendance et de normalité des résidus.

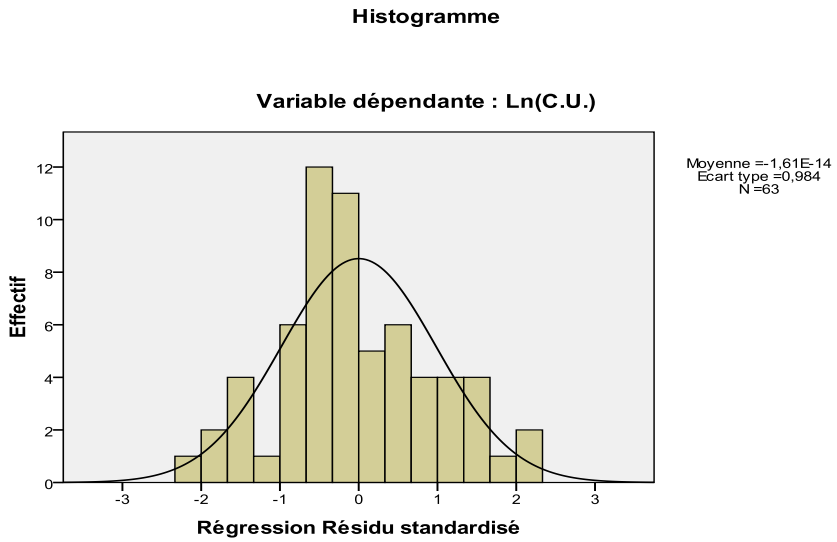

Figure IV-14 : Histogramme des résidus modèle C'

L'histogramme, ainsi que le tableau de comparaison des moyennes montrent l'amélioration des résultats de la simulation :

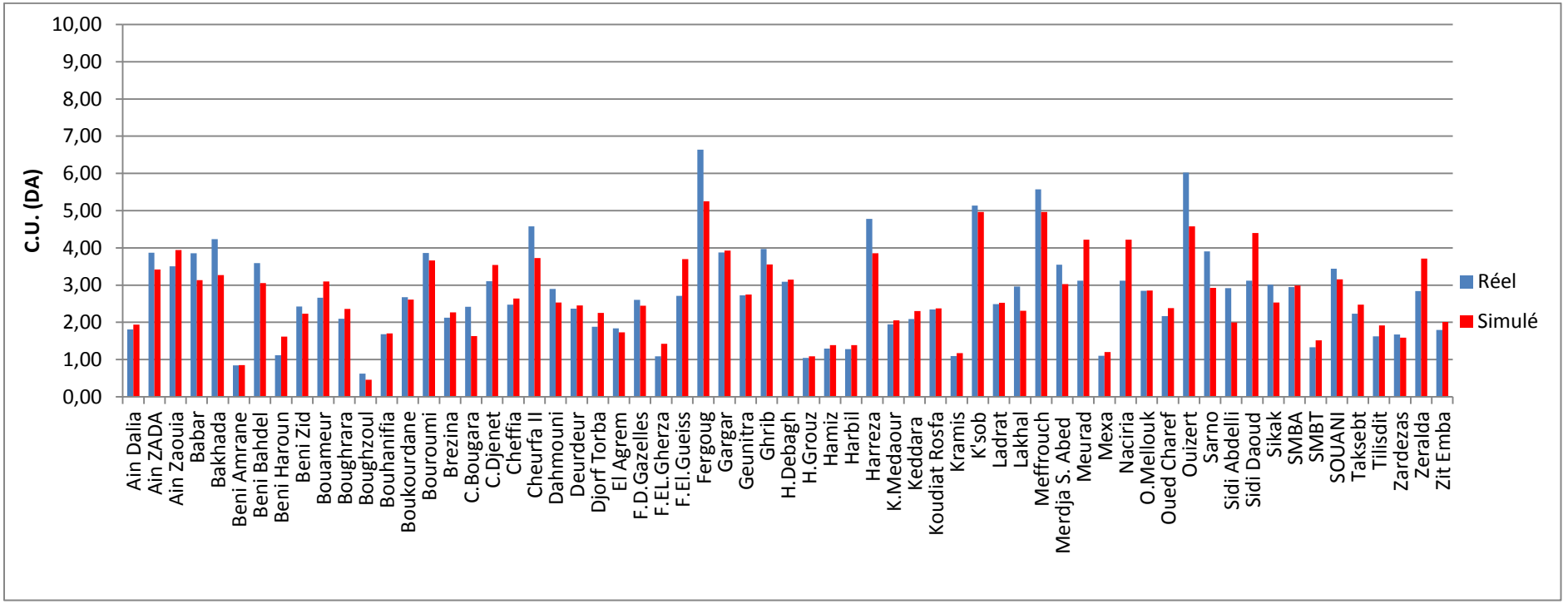

Figure IV-15 : Résultats simulation modèle C'

Tableau IV-21 : Coût de revient unitaire moyen modèle C'

	$C.U._{Moy}$ réel (DA)	$C.U._{Moy}$ Simulé (DA)
Oued Fodda exclu	2,76	2,70
Oued Fodda réintroduit	3,19	3,13

La moyenne simulée se rapproche considérablement de la moyenne calculée, confirmant ainsi qu'une grande part de l'erreur observée dans les résultats du modèle précédent est attribuée au barrage de Oued Fodda. Le modèle C'est donc un bon modèle et peut être utilisé dans l'estimation du coût unitaire moyen.

IV.4.2.6 Modèle D : C.U. = b_0 (Capa) ^{b1} (Vol.mob) ^{b2}

Dernière étape de simplification du modèle de simulation du coût unitaire, nous n'allons garder que les variables les plus faciles à définir, surtout concernant une estimation du coût de revient unitaire standard (dans le but d'évaluer le budget prévisionnel et à fortiori la sujétion de service public).

Ce modèle, n'incluant que les capacités et les volumes mobilisés comme variables explicatives a la forme suivante :

$$C.U. = 10,29 * Capa^{0,510} * Vol.Mob^{-0,673}$$

IV.4.2.6.1 Evaluation du modèle

Tableau IV-22 : Critères d'évaluation modèle D

R^2	R^2 ajusté	F	D-W
0,835	0,829	153,990	2,105

Les critères d'évaluation du modèle ont des valeurs satisfaisantes, les résidus sont indépendants et suivent une loi normale. L'avantage de ce dernier modèle est qu'il ne compte que les volumes comme variables indépendantes, excluant complètement les charges quelles que soient leur nature.

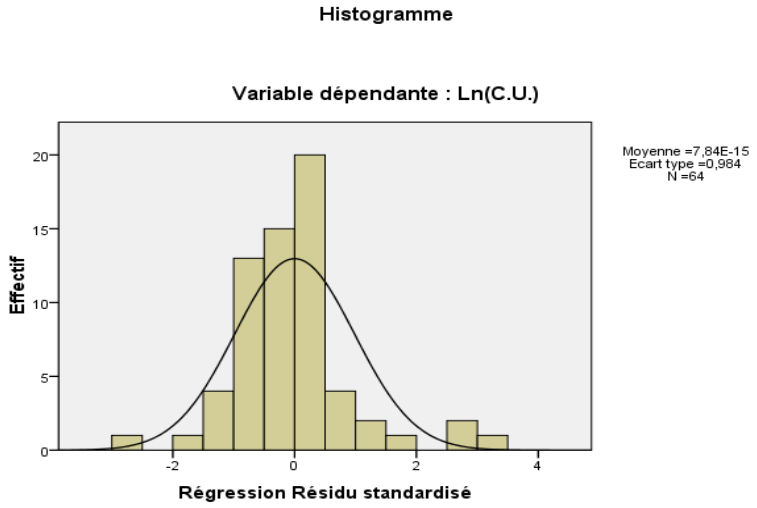

Figure IV-16 : Histogramme des résidus modèle D

Une comparaison entre les coût simulés et réels permet de confirmer cette conclusion.

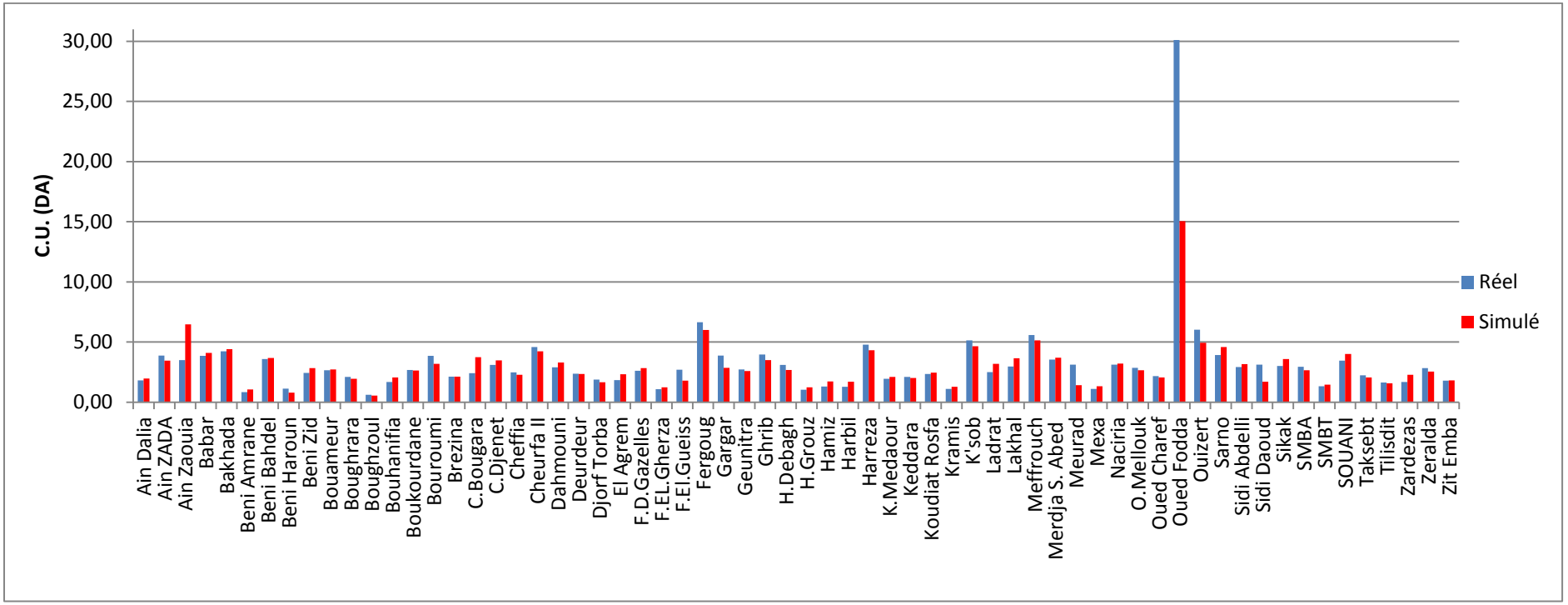

Figure IV-17 : Résultats simulation modèle D

Tableau IV-23 : Coût de revient unitaire moyen modèle D

$C.U._{Moy}$ réel (DA)	$C.U._{Moy}$ Simulé (DA)
3,19	2,97

Le coût unitaire du barrage de Oued Fodda étant toujours largement sous estimé, comme pour les modèles B et C, nous devons l'exclure de l'analyse avant de tirer les conclusions définitives sur les résultats de la simulation.

IV.4.2.7 Modèle D'

Comme pour les modèles B et C, ce modèle est établi afin de réduire l'écart entre les moyennes simulée et réelle causée par la grande sous estimation du coût de revient unitaire du barrage de Oued Fodda (15 DA de différence)

Le modèle de régression établi est

$$C.U. = 9{,}934 * Capa^{0{,}460} * Vol.Mob^{-0{,}614}$$

IV.4.2.7.1 Evaluation du modèle

Tableau IV-24 : Critères d'évaluation modèle D'

R^2	R^2 ajusté	F	D-W
0,807	0,801	125,434	2,205

On remarque une légère baisse de la qualité de l'ajustement, qui reste toutefois satisfaisante. La normalité et l'indépendance des résidus sont quant à elles toujours vérifiées par une valeur de Durbin Watson proche de 2 et l'histogramme des résidus que voici :

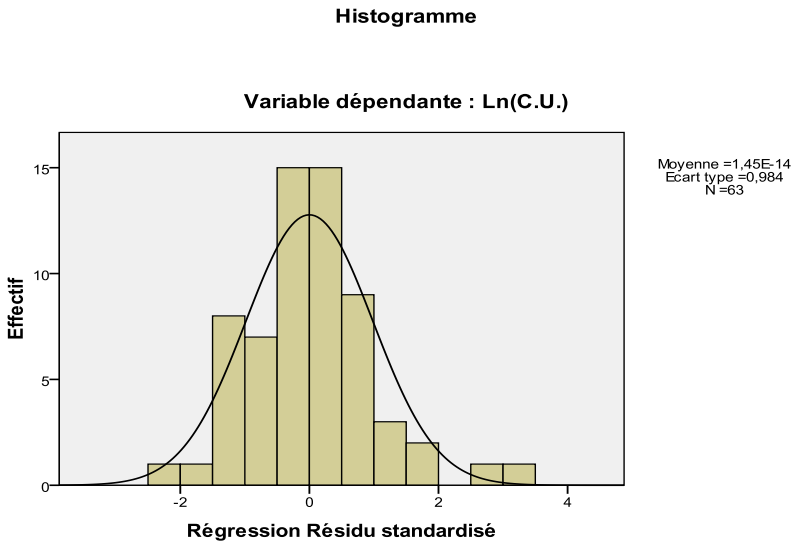

Figure IV-18 : Histogramme des résidus modèle D'

Par contre, le coût de revient unitaire moyen simulé se rapproche considérablement de celui calculé

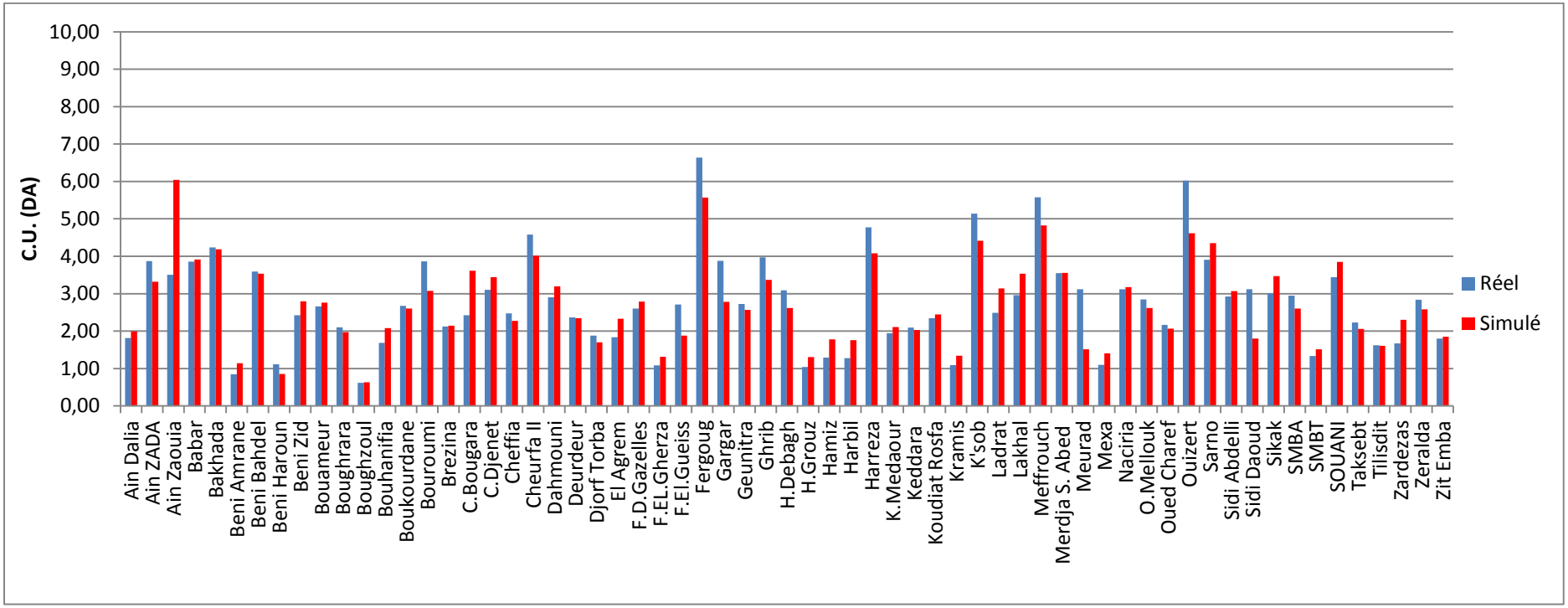

Figure IV-19 : Résultats simulation modèle D'

Tableau IV-25 : Coût de revient unitaire moyen modèle D'

	$C.U._{Moy}$ réel (DA)	$C.U._{Moy}$ Simulé (DA)
Oued Fodda exclu	2,76	2,72
Oued Fodda réintroduit	3,19	3,15

Même s'il existe de légères différences entres les valeurs simulées et calculées, amplifiées visuellement par l'échelle des ordonnées qui n'est pas la même que celle du modèle D, ces différences se compensent et aboutissent à une moyenne simulée très proche de la moyenne réelle.

IV.4.3 Conclusions de la régression

1) Un modèle de régression linéaire multiple classique n'est pas envisageable, il faut passer par une transformation des variables qui est dans notre cas, une transformation logarithmique.
2) Le barrage de Oued Fodda est un barrage atypique responsable à lui seul d'une grande part de l'erreur commise lors de la simulation.
3) Trois modèles (B', C', D') ont été établis. Ces modèles donnent de très bons résultats concernant le coût de revient unitaire moyen. Le choix du modèle utilisé se fera en fonction de la disponibilité des données.
4) Le modèles D', sous réserve d'une validation avec des données actualisées, est le modèle le plus avantageux car il n'a besoin que des capacités et des volumes mobilisés pour simuler le coût unitaire, ces variables étant plus faciles à déterminer par rapport aux charges, surtout dans le but d'une évaluation du coût standard.

Conclusion générale

L'objectif de ce travail était d'établir une démarche, permettant d'estimer le coût unitaire standard dans l'optique d'évaluer la sujétion de service public et déterminer par conséquent le budget prévisionnel qui devra être alloué à l'ANBT pour couvrir toutes les charges auxquelles l'agence doit faire face dans sa mission d'exploitation des barrages.

L'analyse en composantes principales et la régression multiple sont les deux méthodes utilisées dans ce travail.

Bien qu'elle n'ait pas été utilisée de façon directe dans l'établissement des modèles, l'analyse en composantes principales représente un premier traitement des données brutes préalable à l'application de la régression multiple.

Elle nous a permis de tirer quelques conclusions très utiles pour la suite de l'étude, concernant les relations existant entre les variables et au sein de l'ensemble des individus. Grâce à l'ACP nous avons pu constater la redondance contenue dans les variables explicatives, ainsi que les barrages atypiques qui risquent de fausser les résultats de la régression qui sera faite par la suite.

La régression multiple nous a permis d'établir plusieurs modèles permettant de simuler le coût de revient unitaire. Les modèles de la forme $Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots b_p X_p + \varepsilon$ ont donné de mauvais résultats, un modèle global étant impossible à définir de cette manière.

Sept modèles de la forme $Y = b_0 X_1^{b1} X_2^{b2} \ldots X_p^{bp} + \varepsilon$ ont quant à eux, donné d'excellents résultats, ils se distinguent les uns des autres par le nombre et la nature des variables explicatives considérées.

Les modèles B', C' et D' sont les modèles ayant donné les résultats les plus satisfaisants, ils présentent le meilleur compromis entre la précision de la simulation et la nature des variables explicatives qu'ils comprennent.

Ces trois modèles écartent le barrage de Oued Fodda de l'analyse car il est responsable à lui seul d'une grande part de l'erreur constatée dans les résultats des modèles B, C et D.

Nous recommandons l'utilisation du modèle D' dans l'estimation du coût unitaire standard moyen car c'est le modèle comportant les variables indépendantes les plus facilement « mesurables » et de ce fait le plus avantageux des modèles établis dans cette étude.

Il faut toutefois noter que l'absence de données autres que celles de l'exercice 2005/2006 ne permet pas de confirmer la bonne qualité des modèles de régression obtenus. De plus, il est impossible d'évaluer la sujétion de service public faute de disponibilité des données actualisées ou d'estimations des données concernant les années à venir.

Références bibliographiques

Ambapour. S.(2003). *Introduction à l'analyse des données*. Bureau d'Application des Méthodes Statistiques et Informatiques. http://www.cnsee.org/

Amini. M-R. *Techniques d'analyse de données et théorie de l'information*. Laboratoire d'informatique de Paris6. http://www.connex.lip6.fr/

Baccini. A.(2010). *Statistique Descriptive Multidimensionnelle*. Institut de Mathématiques de Toulouse. http://www.math.univ-toulouse.fr/

Baillargeon. J.(2003). *L'analyse en composantes principales*. http://www.uqtr.ca/

Berger. E. D. *Introduction to multiple regression*. Clermont University. http://www.cgu.edu/

Brocard. D.(2010). *Régression multiple*. Université de Montréa. http://www.cgu.edu/

Contribution pour la mission de sujétion de service public : Exposé des motifs. ANBT. (2008)

De Rongé. Y.*Comptabilité analytique et budgétaire*. http://tuyaux.aglouvain.be

Dodge. Y, Rousson. V.(2004) *Analyse de régression appliquée*. Dunod..

Duby. C, Robin. S.(2006). *Analyse en Composantes Principales*. Institut National Agronomique Paris. http://www.agroparistech.fr

Durand. C.(2003) *L'analyse factorielle et l'analyse de fidélité*. Université de Montréal. http://www.mapageweb.umontreal.ca

Engel.F et Kletz.F. *Cours de comptabilité analytique. Ecole des mines de Paris*, 2003.

Gujarati. N. D.(1995) *Basic econometrics, Third Ed*. McGraw Hill.

Laffly. D. *Régression multiple : Principes et exemples d'application*. Université de Pau et des pays de l'Adour. 2006. http://web.univ-pau.fr

Hamriche. S, Tachet. A.(1993).*Contribution à la modélisation des phénomènes hydrométéorologiques par l'analyse en composantes principales*. Ecole Nationale Polytechnique.

Kasmi. M.(2007).T *Manuel des analyses socio-économiques, Vol 3 : Coûts de l'eau des ouvrages hydrauliques en milieu rural*. Bischöfliches Hilfswerk MISEREOR e. V.

Morice. H.(2008).Séminaire généraliste : *Appui méthodologique pour l'élaboration et l'évaluation économique et financière des projets d'aménagement hydrauliques*.

Morineau. A. *ACP-Analyse en Composantes Principales*. Deenov.com

Rakotomalala. R. *Régression linéaire multiple*. Equipe de recherche en Ingénierie des connaissances.

Raufaste. E.(2009). *Techniques de recodage des données*. UOH. http://w3.uohpsy.univ-tlse2.fr