الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

وزارة التعليم العالي والبحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية الرياضيات
Faculté des Sciences

قسم الرياضيات
Département de Mathématiques

Mémoire de Master

Spécialité : Mathématiques

Option : Modélisation Stochastique et Statistique

THEME

Etude et implémentation de la Méthode d'Analyse en Composantes Principales et l'Analyse Factorielle des Correspondances pour Traitement des Données. Application à SONATRACH

Présenté par

BOUABDALLAH M'hamed Akram

&

SAIDI Mohand Said

Devant le jury compose de :

Mr. MASSIED M.	Maître assistant A	Président
Mlle. KERDJOUDJ S.	Maître de conférences A	Examinateur
Mr. ELMOOSAOUI H.	Maître de conférences A	Rapporteur
Mr. SOKHAL A.		Co-Rapporteur

ملخص

في هذا العمل، قدمنا مراجعة شاملة لتحليل البيانات، مع التركيز على تحليل المكونات الرئيسية (ACP) وتحليل التحليل العاملي للمراسلات (AFC). تتكون هذه الطرق من مجموعة من التقنيات التي تساعد في اكتشاف البنية المعقدة لجدول بيانات متعدد الأبعاد وترجمتها بشكل أبسط وملخص، وعادة ما تُمثل بشكل بياني. لقد قمنا بتطوير تطبيق بلغة PYTHON لتلبية احتياجات المستخدمين والدراسات الخاصة.

من خلال تدريبنا العملي في قسم هندسة وتطوير البترول في شركة سوناطراك، تمكنا من تطبيق معارفنا النظرية واستشفاف مزايا تحليل المكونات الرئيسية في إعادة إعمار السجلات. يوفر هذا العمل فهمًا عميقًا للمكونات الرئيسية وفائدتها في صناعة النفط والغاز.

الكلمات الرئيسية: تحليل البيانات، تحليل عاملي، تحليل المكونات الرئيسية، تحليل المراسلات، صناعة النفط والغاز.

Résumé

Dans ce travail, nous avons présenté un état de l'art sur l'analyse des données factorielles, en mettant l'accent sur l'Analyse en Composantes Principales (ACP) et l'Analyse Factorielle des Correspondances (AFC). Ces méthodes regroupent un ensemble de techniques permettant de découvrir la structure complexe d'un tableau de données multidimensionnel et de la traduire de manière plus simple et résumée, souvent représentée graphiquement. Nous avons développé une application en langage PYTHON pour répondre aux besoins des utilisateurs et des études spécifiques.

Grâce à notre stage pratique au sein de la Division Petroleum Engineering & Development de la SONATRACH, nous avons pu appliquer nos connaissances théoriques et constater les avantages de l'ACP dans la reconstitution des logs. Ce travail offre une compréhension approfondie de l'ACP et de son utilité dans le domaine pétrolier.

Mots clés : Analyse des données, Analyse factorielle, Analyse en composantes principales, Analyse des correspondances, Industrie pétrolière et gazière.

Abstract

In this work, we have presented a state-of-the-art review on factorial data analysis, focusing on Principal Component Analysis (PCA) and Correspondence Analysis (CA). These methods encompass a set of techniques that allow discovering the complex structure of multidimensional data tables and translating it into a simpler and summarized form, often represented graphically. We have developed a PYTHON application to cater to the needs of users and specific studies.

Through our practical internship at the Petroleum Engineering & Development Division of SONATRACH, we were able to apply our theoretical knowledge and observe the advantages of PCA in log reconstruction. This work provides an in-depth understanding of PCA and its usefulness in the petroleum industry.

Keywords: Data analysis, Factorial analysis, Principal Component Analysis, Correspondence analysis, Oil and gas industry.

REMERCIEMENTS

Nous remercions en premier lieu **Dieu**, le tout puissant de nous avoir donné tant de courage, de patience et de volonté pour l'élaboration de ce modeste travail.

Nous tenons à remercier Mr Elmossaoui Hichem, Maitre de Conférences à l'Université Blida1, notre directeur de mémoire, qui nous a encadré et conseillé judicieusement durant ce travail.

Nous tenons à remercier Monsieur SOKHAL Abdallah, notre encadreur de stage, pour son soutien constant, ses conseils précieux et son expertise tout au long de cette expérience professionnelle.

Nous exprimons notre profonde gratitude et nos sincères remerciements à Mr O. Tami, le Chef de département de Mathématiques à l'Université Blida1 pour toute l'aide qu'il nous donne.

Nous profitons de cette occurrence pour exprimer notre profonde gratitude et nos sincères remerciements à tous les enseignants qui ont contribué à notre formation.

Nous n'oublierons pas de remercier les membres de la scolarité du département de Mathématiques en particulier : Mr H. hadj allah, Mme N. Djenas et Mme S. Takarli

Enfin, que les membres de jury trouvent ici l'expression de toute notre gratitude pour l'honneur qu'ils nous font en acceptant de juger ce travail

Dédicaces

Je dédie humblement ce travail à mes chers parents qui ont été une source inépuisable de soutien, de patience et d'encouragement tout au long de mon parcours scolaire. Leur amour, leur confiance, leurs conseils ainsi que leur soutien ont été les piliers de ma réussite. Je souhaite également dédier ce travail à ma sœur, mes frères et à toute ma famille, dont le soutien constant et les encouragements m'ont donné la force de persévérer. Enfin, je tiens à exprimer ma reconnaissance à tous mes amis qui ont été présents à mes côtés, partageant des moments précieux et me soutenant tout au long de cette aventure

SAIDI Mohand Said

je dédie ce modeste travail à mes chers parents ,à ma sœur la prunelle de mes yeux et son mari, à mes grands-parents, à mes tantes et oncles, cousins et cousines paternelles et maternelles, mes amies et tous ceux qui m'ont aidé tout au long de mon parcours
Je dédie ce travail spécialement aux êtres chères que j'ai perdu récemment, babasido, amty ouda et amty zehour que dieu leurs accorde le paradis.

BOUABDALLAH M'HAMED AKRAM

Table Des Matières

Liste Des Illustrations, Graphiques Et Tableaux

Introduction

Le présent mémoire constitue une étude consacrée à l'analyse des données. Ce domaine de recherche joue un rôle essentiel dans de nombreux domaines tels que la statistique, la science des données, la sociologie, le marketing, et bien d'autres encore. L'analyse des données vise à extraire des informations utiles et significatives à partir de jeux de données, en utilisant des méthodes et des techniques spécifiques.

La méthode traditionnelle de statistique nous a habitués à étudier les variables individuellement, en construisant des histogrammes pour chacune d'entre elles. Cependant, comment pouvons-nous remplacer ces multiples graphiques par un seul graphique, une carte plane ? Comment pouvons-nous obtenir une vision d'ensemble des résultats plutôt que de se perdre dans une multitude de descriptions partielles résultant de l'analyse variable par variable ? C'est là que les techniques d'analyse des données interviennent pour répondre à ces questions.

L'analyse des données regroupe un ensemble de techniques visant à découvrir la structure d'un tableau de nombres multidimensionnel, tout en cherchant à la traduire par une structure plus simple qui résume au mieux les informations. Cette structure est souvent représentée graphiquement pour une meilleure visualisation.

Ces techniques, principalement descriptives, ont pour objectif de décrire, réduire, classer et clarifier les données en prenant en compte différents points de vue. Elles permettent d'étudier les grandes tendances, les relations, les similarités ou les différences entre les variables ou groupes de variables. Cette approche descriptive et multidimensionnelle peut être considérée comme une forme perfectionnée de la statistique descriptive.

L'analyse des données englobe principalement deux ensembles de techniques. Les premières, relevant de la géométrie euclidienne et basées sur l'extraction de valeurs et de vecteurs propres, sont connues sous le nom d'"analyses factorielles". Les secondes, appelées "classification automatique", se caractérisent par l'utilisation d'un indice de proximité et d'un

algorithme d'agrégation ou de désagrégation, qui permettent d'obtenir une partition ou un arbre de classification.

Notre travail a pour premier objectif de nous concentrer principalement sur les analyses factorielles, en décrivant en détail les deux méthodes les plus couramment utilisées : l'analyse en composantes principales (ACP) et l'analyse factorielle des correspondances (AFC). De plus, nous développons une application en langage PYTHON qui permettra de traiter des tableaux de données et d'afficher graphiquement tous les résultats liés à ces deux méthodes.

Le deuxième objectif de ce travail est de réaliser un stage pratique au sein de la Division Petroleum Engineering & Development de la SONATRACH. L'objectif spécifique est de résoudre les problèmes liés à la qualité des interprétations des anciens puits du champ de MESDAR, qui sont liés à la digitalisation des logs et à l'état du trou. Pour cela, nous utiliserons la méthode de l'Analyse en Composantes Principales (ACP) pour reconstituer les logs affectés. Deux solutions ont été envisagées pour résoudre ce problème. La première solution consistait à faire confiance aux données carotte au détriment des logs, en appliquant un lissage des données carotte (Porosity Windowing). Cependant, la deuxième solution, qui a été retenue, était de reconstituer les logs affectés en utilisant la méthode ACP (Analyse en Composante Principales).

Le mémoire est structuré en quatre chapitres :

- Dans le premier chapitre nous examinerons les différentes formes de données auxquelles nous pouvons être confrontés, qu'elles soient numériques, catégorielles, ou encore temporelles. Nous étudierons également les concepts fondamentaux tels que les variables, les mesures de tendance centrale, de dispersion, et de corrélation, qui sont essentiels pour comprendre les données et leur interprétation.
- Dans le deuxième chapitre, nous nous pencherons sur l'Analyse en Composantes Principales (ACP). Cette méthode statistique puissante permet de réduire la dimension des données en identifiant les variables les plus significatives et en les représentant dans un espace de dimensions réduit.
- Le troisième chapitre sera consacré à l'Analyse Factorielle des Correspondances (AFC), une méthode spécifique à l'analyse des données catégorielles. L'AFC permet d'étudier les relations entre deux variables qualitatifs et de mettre en

évidence des associations significatives et nous illustrerons ces concepts à l'aide d'une application concrète.

- Dans le quatrième chapitre, nous présenterons une application de l'ACP dans un contexte spécifique. Nous mettrons comment l'ACP peut être utilisée pour résoudre des problèmes concrets et prendre des décisions éclairées.
- La description de l'Application élaborée en langage PYTHON a été abordée au chapitre cinq.

Enfin, une conclusion termine ce travail, donnant quelques perspectives pour des travaux futurs et en annexe, nous présentons l'ensemble des résultats relatifs à l'application de la méthode ACP pour d'autres puits.

CHAPITRE I

Nature Des Données Et Concepts Fondamentaux

Dans ce chapitre sont synthétisés et résumés les différentes hypothèses entrant en jeu dans l'utilisation de la méthode d'Analyse de Données. Nous fournira les bases nécessaires pour explorer plus en profondeur l'analyse de données dans les chapitres suivants. Nous acquerrons une compréhension solide des données et des concepts fondamentaux qui nous aideront à mener des analyses efficaces et à prendre des décisions éclairées dans divers domaines.

1.1. Un peu d'Historique

Bien que l'étude de la structure des grands ensembles de données soit récente, les méthodes d'analyse de données s'inspirent de principes anciens. L'analyse factorielle trouve ses origines dans les travaux de Ch. Spearman en 1904, qui a introduit le concept de facteur pour expliquer les résultats obtenus par de nombreux sujets dans divers tests. Dans les années 30, des chercheurs tels que C. Burt et L.L Thurstone se sont intéressés à la recherche de plusieurs facteurs, tels que la mémoire et l'intelligence, qui ne sont pas directement observables mais qui peuvent expliquer statistiquement les nombreuses observations. Cependant, de nos jours, l'analyse factorielle au sens des psychologues est moins utilisée car elle suppose un modèle a priori.

Ensuite, l'analyse factorielle en composantes principales a été développée par H. Hotelling en 1933, bien que le principe puisse être attribué à K. Pearson en 1901. Dans cette approche, les individus représentés par les colonnes du tableau sont considérés comme des vecteurs dans un espace multidimensionnel, et l'objectif est de réduire la dimension de l'espace en projetant le nuage de points des individus sur un sous-espace de dimension pk (avec k fixé à une petite valeur), permettant ainsi un ajustement optimal du nuage. Selon L. Lebart, l'analyse en composantes principales est une technique de représentation des données

qui présente un caractère optimal selon certains critères algébriques et géométriques spécifiés, et elle est généralement utilisée sans référence à des hypothèses statistiques spécifiques ou à un modèle particulier [1, 2].

En dernier lieu, l'analyse factorielle des correspondances, introduite par J.P. Benzécri en 1962, est actuellement très répandue. Elle offre des représentations simplifiées dans une certaine mesure, sans faire d'hypothèses a priori, facilitant ainsi leur interprétation. Pour citer le Professeur J.P. Bensécri : "L'analyse des correspondances, telle qu'elle est pratiquée en 1977, ne se limite pas à l'extraction de facteurs à partir de tableaux de nombres positifs. Elle propose des règles de préparation des données telles que le codage sous forme disjonctive complète, aide à critiquer la validité des résultats par des calculs de contribution, fournit des procédures efficaces de discrimination et de régression, et s'intègre harmonieusement à la classification automatique" [3]. Sa logique est claire : le modèle doit suivre les données, et non l'inverse, le modèle probabiliste est considéré comme trop contraignant : "la statistique n'est pas la probabilité".

Pour terminer cette page d'histoire, mentionnons l'analyse des données non métriques introduite par une nouvelle école de statisticiens américains sous le nom de " multidimensional scaling" (J.D. Carrol, J.B. Kruskal, R.N. Shepard, …) et dont les principales méthodes sont : l'analyse des proximités, l'analyse des préférences et l'analyse de mesure conjointe (qui permet d'expliquer une variable qualitative ordinale à l'aide des variables nominales). Ces méthodes ont trouvé leurs applications surtout dans le domaine du marketing [4].

1. 2. Qu'est-ce que l'Analyse des données ?

L'analyse de données englobe différentes techniques statistiques utilisées pour recueillir, organiser, présenter et étudier des données. Son objectif est de permettre aux experts de tirer des conclusions et de prendre des décisions éclairées. La première étape de cette analyse consiste à définir la population ou les individus à étudier. Ces individus sont caractérisés par des variables ou des caractères. Généralement, ces individus et variables sont présentés sous forme de tableau ou de matrice.

La récolte et l'étude des données sont importantes depuis longtemps pour la survie de plusieurs activités professionnelle et la réputation de différents organismes. Les médecins

établissent des diagnostics en analysant les données de leurs patients. Les entreprises maintiennent ou changent leurs décisions selon les données des marchés et l'appréciation des consommateurs. Les développeurs collectent et traitent les données sur la qualité et la crédibilité de leurs produits, …etc.

L'analyse des données est un processus universel d'acquisition de connaissances, elle requiert des méthodes qui lui permettent d'interpréter avec rigueur les données numériques qu'elle utilise [5, 6].

1.3. Description des données

La description des données est une étape essentielle de l'analyse statistique. Elle vise à présenter de manière concise et précise les caractéristiques principales des données collectées. Cette description inclut des informations sur la nature des variables, leur échelle de mesure, ainsi que des résumés statistiques tels que la moyenne, la médiane et l'écart-type. Elle permet aux analystes de mieux comprendre les données avant de les explorer plus en profondeur et d'en tirer des conclusions pertinentes

1.3.1. Données et leurs caractéristiques

1.3.1.1. Individus et variables

Les individus et les variables sont définit ci-dessous.

Définition 1.1 (Individu) : Le i^{éme} individu est un vecteur à p composantes réelles qu'on le note par e_i tel que :

$$\boldsymbol{e_i = (x_{i1}, x_{i2}, \ldots\ldots, x_{ip})^t \epsilon\, \mathbb{R}^p, pour\ i = \overline{1, p}}$$

Définition 1.2 (Variable) : La j^{éme} variable est la liste des n valeurs qui elle prend sur n individus, on la note par x_j tel que :

$$\boldsymbol{x_j = (x_{1j}, x_{2j}, \ldots\ldots, x_{nj})^t \epsilon\, \mathbb{R}^n, pour\ j = \overline{1, n}}$$

Dans notre contexte, nous traitons principalement de variables quantitatives, qui se distinguent des variables qualitatives.

Définition 1.3 (Variable quantitative) : En statistique, une variable quantitative est une variable qui reflète une notion de grandeur c'est-à-dire si les valeurs qu'elle peut prendre sont des nombres. Une grandeur quantitative est souvent exprimée avec une unité de mesurequi sert de référence.

Définition 1.4 (Variable qualitative) : En statistique, une variable qualitative est une variable catégorielle (facteur) qui prend pour valeur des modalités (catégories, niveaux, …etc.), par opposition aux variables quantitatives qui mesurent sur chaque individu une quantité.

Il faut noter que, la distinction entre ces deux types de variables est importante dans l'analyse des données, car elles nécessitent des approches et des techniques différentes pour leur traitement et leur interprétation.

1.3.1.2. Les tableaux de données

On distingue généralement deux ensembles : les individus et les caractères relatifs à ces individus. Les caractères observés peuvent être quantitatifs s'ils prennent des valeurs numériques. Ils sont dits qualitatifs lorsqu'ils possèdent des modalités non numériques.

a. Tableaux individus-caractères quantitatifs

Les p caractères quantitatifs $x^1, x^2, \dots, x^p$ sont observés sur un ensemble de n individus $e_1, e_2, \dots, e_n$. On obtient une matrice Individus-Caractères que l'on notera $X = (x_{ij})$ et dont la dimension est $(n \times p)$ (n lignes et p colonnes).

$$X = \begin{array}{c} \\ e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{array} \begin{array}{c} \begin{array}{cccccc} x^1 & \cdots & x^j & \cdots & x^p \end{array} \\ \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} \end{array}$$

b. Tableaux individus - caractères qualitatifs

Sur les mêmes individus, on aurait pu observer les caractères qualitatifs : sexe, niveau hiérarchique, situation matrimoniale. Ces caractères peuvent avoir plusieurs modalités. Pour le traitement numérique, ces caractères qualitatifs sont représentés sous forme d'un tableau

de variables indicatrices prenant les valeurs 0 ou 1. On dit alors que les données sont représentées sous forme disjonctive. Cette représentation des caractères qualitatifs permet de les assimiler à des caractères quantitatifs prenant la valeur 0 et 1. On verra également sa fécondité puisque tout tableau de données contenant simultanément des caractères quantitatifs et qualitatifs peut être représenté ainsi. En effet un caractère quantitatif peut être rendu qualitatif par découpage en classes de ses valeurs, puis représenté sous forme de variables indicatrices (classes de revenus, classes d'âges… etc.) [7].

Tableau 1.1 : Tableau individus-caractères qualitatifs.

Caractères

Individus	**Sexe**		**Situation Matrimoniale**			**Niveaux hiérarchique**		
	Masculin	**Féminin**	**Marié**	**Célibataire**	**Veuf**	**Cadre**	**Ouvrier**	**Maitrise**
	1	0	1	0	0	1	0	0
	0	1	0	1	0	0	0	1
	…	…	...	...	...	...	…	…
	1	0	0	0	1	0	1	0
	…	…	...	...	...	…	…	…
	0	1	0	1	0	1	0	0
	1	0	1	0	0	0	0	1

c. Tableaux de Contingence

Un tableau de contingence ou tableau croisé contient les effectifs ou fréquences d'association entre les modalités de deux caractères qualitatifs observés sur un ensemble de n individus. On peut par exemple considérer le tableau croisé des catégories socioprofessionnelles (p modalités) avec les quartiers d'une ville (q modalités). Une case (i, j) de ce tableau contient le nombre n_{ij} d'individus exerçant la Profession i et habitant le Quartier j. Ce tableau sera noté $N = (n_{ij}),\ i = 1, \dots, p$ et $j = 1, \dots, q$.

Tableau 1.2 : Tableau de Contingence

$$
\begin{array}{cc}
 & \textit{Caractère 2} \\
 & \begin{array}{cccccc} \textit{modalité 1} & \cdots & \textit{modalité } j & \cdots & \textit{modalité } q \end{array} \\
\textit{Caractère 1} & \begin{array}{l} \textit{modalité 1} \\ \vdots \\ \textit{modalité } i \\ \vdots \\ \textit{modalité } p \end{array}
\begin{bmatrix} n_{11} & & \cdots & & n_{1q} \\ \vdots & & \vdots & & \\ n_{i1} & & n_{ij} & & n_{iq} \\ \vdots & & \vdots & & \\ n_{p1} & & \cdots & & n_{pq} \end{bmatrix}
\end{array}
$$

Dans un tel tableau, les individus ont été regroupés et ne peuvent plus être distingués. On peut concevoir une autre représentation en dissociant les deux caractères. A chacun d'eux on associe un tableau de variables indicatrices X_1 et X_2. En ligne on représente les individus et en colonne les modalités de chaque caractère. Une ligne ne contient alors que des 0 sauf dans les colonnes correspondant respectivement à la profession ou au quartier considérée où l'on trouve des 1. Notons que le tableau de contingence N est le résultat du produit matriciel $X^t{}_1 X_2$ où ${X_1}^t$ désigne la matrice transposée de la matrice X_1.

$$N = {X_1}^t X_2$$

Par exemples considérons huit individus possédant trois fonctions et habitant deux quartiers :

Profession Quartiers

$$X_1 = \begin{matrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \end{matrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad X_2 = \begin{matrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \end{matrix}\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

Nous aurons donc le tableau de contingence suivant :

Quartier

$$N = {X_1}^t X_2 = \text{ Profession } \begin{bmatrix} 2 & 0 \\ 2 & 1 \\ 1 & 2 \end{bmatrix}$$

d. Tableaux de proximité

Etant donné un ensemble d'individus. On dispose d'une mesure de ressemblance ou de dissemblance entre ces individus pris deux à deux. On peut alors construire un tableau de proximité dans lequel les lignes ainsi que les colonnes représentent ces individus. Un tel tableau est carré, symétrique et généralement ne contient que des nombres positifs. On peut citer par exemple, le tableau où figurent les distances entre les principales villes d'un pays.

1.3.2. Matrice des poids

Si les données ont été recueillies d'un tirage aléatoire, alors les probabilités de choix de ces n individus ont toutes la même importance égale à $\frac{1}{n}$, or ceci n'est pas toujours le cas. Dans le cas contraire, il est utile de travailler avec des poids qu'on note par p_i pour les différents individus où ces derniers sont regroupés dans une matrice diagonale de taille n notée D et appelée matrice des poids. Elle est définie comme suit :

$$D = \begin{bmatrix} P_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_n \end{bmatrix},$$

Avec, $0 \leq p_i \geq 1$ et $\sum_{i=1}^{n} p_i = 1$.

Proposition 1.1 : Dans le cas usuel des poids égaux, nous avons : $D = \frac{1}{n} I_n$.

Preuve : Comme on a $p_1 = p_2 = ...p_i = ... = p_n$ et $\sum_{i=1}^{n} p_i = 1$ *alors :* $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} p_1 = p_1 \sum_{i=1}^{n} 1 = p_1 n = 1$. Par conséquent $p_1 = p_i = \frac{1}{n}$. Et, on aura :

$$D=\begin{bmatrix} \frac{1}{n} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{n} \end{bmatrix} = \frac{1}{n}\begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} = \frac{1}{n} I_n. \text{ Ou } 1_n = vecteur\ unitaire$$

1.3.3. Centre de gravité

Le vecteur des moyennes arithmétiques de chaque variable, noté g et également appelé individu moyen ou point moyen, est défini comme suit :

$$\boldsymbol{g = (\overline{x_1}, \overline{x_2}, \ldots, \overline{x_p})^t \epsilon\ \mathbb{R}^p},$$

Avec, $\boldsymbol{\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\ , j = 1 \ldots p}$.

La forme matricielle de cette définition est la suivante :

$$g = X^t D 1_n. \text{ Ou } 1_n = vecteur\ unitaire$$

Preuve :

Ensuite, vous pouvez procéder à la démonstration en effectuant le remplacement des symboles par leurs valeurs numériques ou en développant les opérations matricielles. Cela permettra de visualiser et de comprendre plus en détail comment la formule fonctionne.

$$\boldsymbol{g = X'D1_n} = \begin{bmatrix} \boldsymbol{x_{11}} & \cdots & \boldsymbol{x_{n1}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{x_{1p}} & \cdots & \boldsymbol{x_{np}} \end{bmatrix} \begin{bmatrix} \boldsymbol{p_1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{p_n} \end{bmatrix} \begin{bmatrix} \boldsymbol{1} \\ \vdots \\ \boldsymbol{1} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \boldsymbol{p_i x_{i1}} \\ \vdots \\ \sum_{i=1}^{n} \boldsymbol{p_i x_{ip}} \end{bmatrix} = \begin{bmatrix} \overline{\boldsymbol{x_1}} \\ \vdots \\ \overline{\boldsymbol{x_p}} \end{bmatrix}$$

1.3.3. Matrice de variance-covariance

La matrice de variance-covariance représente l'ensemble des variances et des covariances, organisées dans un tableau noté S avec un terme général :

$$s_{jj'} = cov(x_j, x_{j'}) = \sum_{i=1}^{n} p_i \left(x_{ij} - \overline{x_j}\right)\left(x_{ij'} - \overline{x_{j'}}\right), pour\ j, j' = 1, \ldots, p.$$

La matrice de variance-covariance est définie donc par :

$$\mathrm{S} = \begin{bmatrix} s_1^2 & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_p^2 \end{bmatrix}$$

La forme matricielle de cette définition est la suivante :

$$S = X^t DX - gg^t = Y^t DY.$$

Avec, $Y = X - I_n g^t$. Dans le cas où les poids sont égaux à $\frac{1}{n}$, cette forme matricielle devient :

$$\mathrm{S} = \frac{1}{n} Y^t Y = \frac{1}{n} X^t X - gg^t$$

Par la suite, nous pouvons procéder à la démonstration en substituant les symboles par leurs valeurs numériques où en développant les opérations matricielles.

$$\mathrm{S} = (\mathrm{X} - I_n g^t)^t\ D\ (X - I_n g^t) = X^t DX - X^t DI_n g^t - gI_n^t DX + gI_n^t DI_n g^t$$

$$= \mathrm{X^t\,DX} - \mathrm{X^t\,D}I_n g^t - \mathrm{g}I_n^t DX + \mathrm{g}I_n^t DI_n g^t$$

$= X^t DX - gg^t - gg^t + gg^t$, car $I_n^t D I_n = \sum_{i=1}^{n} p_i = 1$

$= X^t DX - gg^t$.

1.3.4. Matrice de corrélation

La matrice de corrélation est constituée des coefficients de corrélation, regroupés dans un tableau noté R, où les termes diagonaux sont égaux à 1. Chaque élément $r_{jj'}$ est défini par :

$$r_{jj'} = \frac{s_{jj'}}{s_j s_{j'}}$$

La matrice de corrélation est donnée par :

$$R = \begin{bmatrix} 1 & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & 1 \end{bmatrix}$$

Elle peut également être représentée sous forme matricielle :

$$R = D_{1/s}\, S\, D_{1/s} = Z^t DZ.$$

Avec, $Z = Y\, D_{1/s}$

Ensuite, nous pouvons procéder à la démonstration en remplaçant les symboles par leurs valeurs numériques où en effectuant les opérations matricielles nécessaires pour développer la formule. Cela nous permettra d'obtenir une expression concrète et de mieux comprendre le fonctionnement de la matrice de corrélation.

$$R = D_{1/s}\, S D_{1/s} = \begin{bmatrix} 1/s_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/s_p \end{bmatrix} \begin{bmatrix} s_1^2 & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_p^2 \end{bmatrix} \begin{bmatrix} 1/s_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/s_p \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \cdots & s_{1p}/s_1 s_p \\ \vdots & \ddots & \vdots \\ s_{p1}/s_{p1} s_1 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} 1 & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & 1 \end{bmatrix}$$

Par la suite, nous pouvons démontrer que $Z^t DZ$ est équivalent à R. Pour cela, nous avons :

$$Z^tDZ = (Y\,D_{1/s})^t\,D\,(Y\,D_{1/s}) = D_{1/s}\,Y^t\,DY\,D_{1/s} = D_{1/s}\,S\,D_{1/s} = R$$

Une remarque importante à souligner est que les matrices R et S sont toutes deux des matrices carrées symétriques d'ordre p. Étant donné qu'il y a p variables, cela nous conduit à calculer $\frac{p(p-1)}{2}$ corrélations distinctes.

1.4. Nuage de points (individus)

Chaque individu est représenté par p coordonnées, ce qui lui confère la forme d'un vecteur dans un espace vectoriel défini dans $\mathbb{R}^p$. Cet espace est appelé l'espace des individus. Lorsque nous considérons l'ensemble des n individus, nous obtenons un nuage de points qui est communément appelé le nuage des individus.

1.4.1 Ressemblance entre deux individus

La similarité entre deux individus est déterminée par la proximité de leurs valeurs pour l'ensemble des variables. Cette similarité peut être exprimée par une distance, calculée comme suit :

$$d^2(e_i, e_{i'}) = \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2, \text{ pour } i, i' = 1, \dots, n$$

1.4.2. Métrique

En physique, la distance entre deux points dans l'espace peut être facilement calculée en utilisant la formule de Pythagore. Le carré de la distance est égal à la somme des carrés des différences des coordonnées, car les dimensions ont la même nature et unité de mesure. Cependant, en statistique, la situation est différente car chaque dimension correspond à une variable qui est exprimée dans son propre système d'unité. On particulier, pour résoudre ce problème on définit la distance entre deux individus e_i et $e_{i'}$ sous la forme quadratique suivante :

$$\langle e_i, e_{i'} \rangle_M = (e_i - e_{i'})^t M (e_i - e_{i'})$$

Avec, M est une matrice carrée symétrique d'ordre p définie positive. Dans ce contexte, la formule de Pythagore peut être réinterprétée pour définir le produit scalaire de deux individus e_i et $e_{i'}$. Cette définition est donnée par la relation suivante :

$$\langle e_i, e_{i'} \rangle_M = e_i^t M e_{i'}$$

Les métriques les plus couramment utilisées sont les métriques diagonales, telles que $I_p$et D_{1/s^2}. La métrique I_p représente la matrice identité d'ordre p et la métrique D_{1/s^2} la matrice suivante :

$$D_{1/s^2} = \begin{bmatrix} 1/s_1^2 & \dots & 0 \\ & \ddots & \\ 0 & \dots & 1/s_p^2 \end{bmatrix}$$

Cela signifie essentiellement que chaque caractère est divisé par son écart-type. Cela présente l'avantage de rendre la distance entre deux individus indépendants des unités de mesure, ce qui est particulièrement utile lorsque les variables sont exprimées dans des unités différentes.

Si nous désignons par $e_i^y = \left(y_{i1}, \dots, y_{ip}\right)^t$ le i^{eme} individu du tableau Y, et par $e_i^z = \left(z_{i1}, \dots, z_{ip}\right)^t \in \mathbb{R}^P$ le i^{eme} individu du tableau Z, alors le calcul du produit scalaire peut s'écrire comme suit :

$$\begin{aligned} \left\langle e_i^y, e_i^y \right\rangle_{D_{1/s^2}} &= \left(e_i^y\right)^t D_{1/s^2} e_i^y \\ &= \left(\frac{y_{i1}}{s_1}\right)^2 + \cdots + \left(\frac{y_{ip}}{s_p}\right)^2 \\ &= \sum_{j=1}^{p} \left(\frac{y_{jj}}{s_j}\right)^2 \\ &= \sum_{j=1}^{p} \left(z_{ij}\right)^2 \\ &= \sum_{j=1}^{p} \left(\frac{z_{ij}}{1}\right)^2 \\ &= (e_i^z)^t I_p e_i^z \\ &= \langle e_i^z, e_i^z \rangle_{I_p} \end{aligned}$$

1.4.3 Inertie

On appelle inertie totale du nuage de points la moyenne des carrés des distances des n points par rapport au centre de gravité g. Elle peut être exprimée de la manière suivante [8] :

$$I_g = \sum_{i=1}^{n} p_i d_M^2(e_i, g)$$

On peut également l'exprimer de la manière suivante :

$$I_g = \sum_{i=1}^{n} p_i \|e_i - g\|_M^2 = \sum_{i=1}^{n} p_i (e_i - g, e_i - g)_M = \sum_{i=1}^{n} p_i (e_i - g)^t M (e_i - g)$$

On définit ainsi l'inertie en un point quelconque a différent du centre de gravité comme suit :

$$I_a = \sum_{i=1}^{n} p_i d_M^2(e_i, a)$$

Si $a =$ g, nous avons :

$$I_g = \sum_{i=1}^{n} p_i \|e_i\|_M^2 = \sum_{i=1}^{n} p_i e_i^t M e_i.$$

I_g est relié à I_a par la formule de Huygens suivante :

$$I_a = I_g + \| g - a \|_M^2$$

Cette inertie peut également être exprimée par la relation suivante :

$$I_g = \text{trace}(MS) = \text{trace}(SM)$$

Montrons la relation de Huygens, pour cela, nous avons :$I_a = \sum_{i=1}^{n} p_i \langle e_i - a, e_i - a \rangle_M$ alors :

$$\begin{aligned}\langle e_i - a, e_i - a\rangle_M &= \langle e_i - g + g - a, e_i - g + g - a\rangle_M \\ &= \langle e_i - g, e_i - g\rangle_M + \langle e_i - g, g - a\rangle_M + \langle g - a, e_i - g\rangle_M + \langle g - a, g - a\rangle_M \\ &= \|e_i - g\|_M^2 + 2\langle e_i - g, g - a\rangle_M + \| g - a \|_M^2\end{aligned}$$

D'où

$$\begin{aligned}I_a &= \sum_{i=1}^{n} p_i\big(\|e_i - g\|_M^2 + 2(g - a, e_i - g\rangle_M + \| g - a \|_M^2\big) \\ &= I_g + 2\sum_{i=1}^{n} p_i(g - a, e_i - g\rangle_M + \| g - a \|_M^2.\end{aligned}$$

Il reste à montrer que $\sum_{i=1}^{n} p_i \langle g - a, e_i - g\rangle_M = 0$. En effet,

$$\begin{aligned}\sum_{i=1}^{n} p_i \langle g - a, e_i - g\rangle_M &= \sum_{i=1}^{n} p_i (g - a)^t M (e_i - g\rangle \\ &= (g - a)^t M \sum_{i=1}^{n} p_i (e_i - g) \\ &= (g - a)^t M \left(\sum_{i=1}^{n} p_i e_i - \sum_{i=1}^{n} p_i g\right) \\ &= (g - a)^t M (g - g), \text{car } g = \sum_{i=1}^{n} p_i e_i \\ &= 0\end{aligned}$$

Montrons maintenant que $I_g = \text{trace}(MS) = \text{trace}(SM)$. Nous avons pour cela, la relation suivante :

$$\begin{aligned}I_g &= \sum_{i=1}^{n} p_i (e_i - g)^t M (e_i - g) \\ &= \text{tr}\left(\sum_{i=1}^{n} p_i (e_i - g)^t M (e_i - g)\right) \\ &= \sum_{i=1}^{n} \text{tr}\big(p_i (e_i - g)^t M (e_i - g)\big)\end{aligned}$$

$$= \sum_{i=1}^{n} \text{tr}(p_i M(e_i - g)(e_i - g)^t), \text{car tr}(AB) = \text{tr}(BA)$$
$$= \text{tr}\left(M\left(\sum_{i=1}^{n} p_i (e_i - g)(e_i - g)^t \right) \right)$$
$$= \text{tr}(MS)$$

Proposition 1.2

1. Si $M = I_p$, l'inertie est egale à la somme des variances des p variables:

$$I_g = \sum_{j=1}^{p} S_j^2.$$

2. Si $M = D_{1/s^2}$, linertie est egale au nombre de variables: $I_g = p$

1. 5. Nuage de points (variables)

Chaque variable x_i est une liste de n valeurs numériques qui peut être considérée comme un vecteur de l'espace $\mathbb{R}^n$ appelé espaces des variables. La métrique utilisée pour le calcul des distances entre variables est la métrique D_p.

Soient les variables $x^1, x^2, \ldots, x^p$ centrées. Voici les propriétés importantes à considérer :

- Le produit scalaire entre deux variables x^k et x^i est

$$< x^k, x^i > = (x^k)^t D_p x^i = S_{ki}$$

- Le carré de la norme d'une variable est égal à sa variance

$$\|x^k\|^2 = S_k^2$$

Et l´écart-type représente donc sa longueur,

- Le cosinus de l'angle θ_{ki} entre deux variables x^k et x^i est leur coefficient de corrélation linéaire :

$$cos(\theta_{ki}) = \frac{<x_k, x_i>}{\|x_k\| \|x_i\|} = \frac{S_{ki}}{S_k S_i}$$

1.6. Outils d'analyse des données

Il existe une variété d'outils d'analyse de données sur le marché, chacun offrant ses propres fonctionnalités. Le choix des outils doit toujours être basé sur le type d'analyse à effectuer et le type de données à traiter. Voici quelques-uns des outils les plus convaincants pour l'analyse des données :

1.6.1. Excel

Il dispose d'un grand nombre de fonctionnalités intéressantes et, avec l'installation de plugins supplémentaires, il peut traiter une quantité massive de données. Ainsi, si vous disposez de données qui sont loin de la marge significative de données, alors Excel peut être un outil très polyvalent pour l'analyse des données.

1.6.2. Tableau

Il entre dans la catégorie des outils de BI, conçus dans le seul but d'analyser les données. L'essence même de Tableau est le tableau croisé dynamique et le graphique croisé dynamique, qui permettent de représenter les données de la manière la plus conviviale possible. Il dispose également d'une fonction de nettoyage des données ainsi que de brillantes fonctions analytiques.

1.6.3. Langage R

Il existe plusieurs langages de programmation puissants et flexibles, parmi lesquels R se distingue particulièrement pour l'analyse statistique. R est largement utilisé pour modéliser des distributions normales, implémenter des algorithmes de classification en grappes et effectuer des analyses de régression. Il permet également de réaliser des analyses prédictives individuelles, telles que l'étude du comportement des clients, l'analyse de leurs dépenses et la recommandation d'articles en fonction de leur historique de navigation. De plus, R intègre des concepts d'apprentissage automatique et d'intelligence artificielle pour des analyses avancées.

1.6.4. Langage Python

Python est un langage de programmation polyvalent et puissant. Sa syntaxe claire et lisible facilite l'écriture du code. Avec son typage dynamique, vous n'avez pas besoin de déclarer le type des variables. Il est interprété, ce qui signifie que le code est exécuté ligne par ligne, offrant une portabilité multiplateforme

1.6.5. SAS

SAS est un langage de programmation spécialement conçu pour l'analyse et la manipulation des données. Il offre une grande flexibilité et permet d'accéder facilement à des données provenant de différentes sources. SAS propose également une gamme complète de produits permettant de profiler les clients, d'analyser les données web, les médias sociaux et le marketing. Avec SAS, vous pouvez prévoir les comportements des clients, gérer et optimiser les communications de manière efficace.

1.7. Conclusion

En conclusion, ce premier chapitre consacré aux concepts fondamentaux de l'analyse de données a permis d'acquérir une compréhension approfondie de ce domaine essentiel. Nous avons exploré les différentes définitions et concepts clés, ainsi que les objectifs et enjeux associés à l'analyse de données.

De plus, nous avons examiné les outils et logiciels d'analyse de données disponibles, ce qui nous a offert une perspective sur les ressources pratiques à notre disposition dans ce domaine. Cette exploration nous a donné une base solide pour aborder les prochains chapitres de cette thèse, au cours desquels nous approfondirons nos connaissances et compétences en matière d'analyse de données.

En résumé, ce chapitre a jeté les bases nécessaires pour notre parcours dans le domaine de l'analyse de données. Il nous a fourni une vision d'ensemble des concepts clés et des outils pratiques, nous préparant ainsi à explorer plus en détail les méthodes et les techniques spécifiques dans les chapitres à venir.

CHAPITRE II

Analyse en composantes principales

Dans ce chapitre nous représentons l'Analyse en Composantes Principales (ACP) une méthode statistique puissante utilisée pour analyser un grand nombre de variables quantitatives simultanément. Son objectif principal est de révéler les relations essentielles entre ces variables et de les représenter graphiquement de manière compréhensible. Grâce à l'ACP, il est possible d'explorer les similarités et les différences entre les individus, ainsi que les associations entre les différentes variables. ACP, c'est un outil précieux pour simplifier et visualiser des ensembles de données complexes, permettant ainsi une meilleure compréhension et interprétation des informations statistiques.

2.1. Principe de la méthode

L'ACP vise à réduire la dimension des données initiales, qui peuvent comporter un grand nombre de variables (p), en les remplaçant par un nombre réduit de facteurs (q) appropriés ($q < p$) appelés composantes principales. Ces facteurs recherchés sont des combinaisons linéaires pondérées des p variables d'origine. Leur sélection est basée sur la maximisation de la variance des individus par rapport à ces facteurs. Pour cela, l'ACP utilise des techniques mathématiques adaptées pour effectuer ce processus de manière automatique et optimale, permettant ainsi de simplifier la représentation des données tout en conservant les informations essentielles. Cette approche permet de visualiser et d'analyser les relations et les structures sous-jacentes des données multidimensionnelles de manière plus concise et significative [9].

2.2. Ajustement dans l'espace des individus $\mathbb{R}^p$

Dans cette partie, nous aborderons la construction du sous-espace $\mathbb{R}^q$, qui comprend le nuage de projection et les droites également appelées axes principaux. Commençons par

chercher un sous espace vectoriel à une dimension, c'est-à-dire une droite passant par l'origine et qui réalise le meilleur ajustement possible du nuage de points :

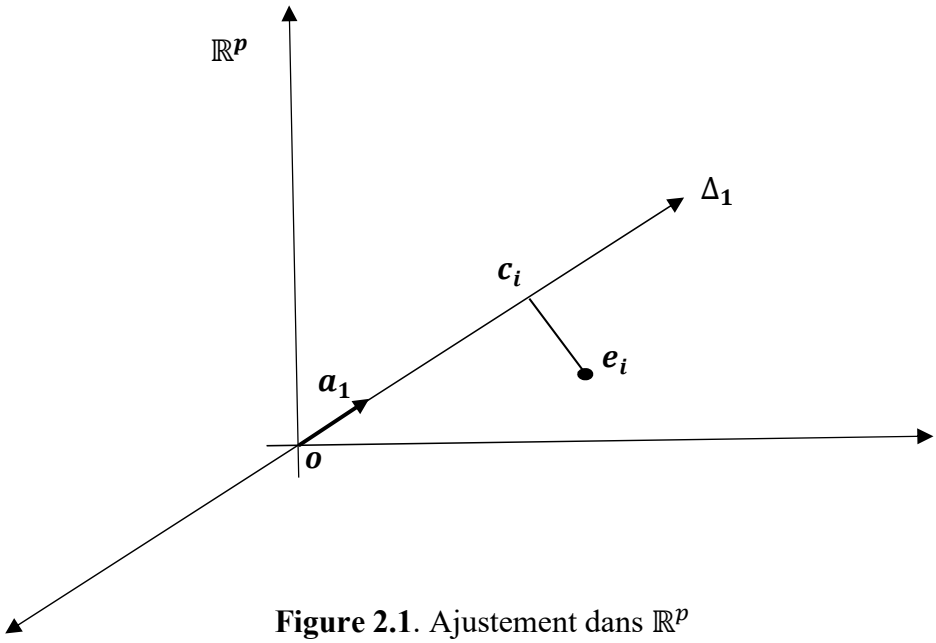

Figure 2.1. Ajustement dans $\mathbb{R}^p$

Le critère du choix de l'espace de projection s'effectue tel que la moyenne des carrés des distances entre les projections et leur centre de gravité soit le plus grand possible. Ce qui implique qu'il faut que l'inertie du nuage projeté sur ce sous-espace soit maximal.

Parmi les critères d'ajustement d'un sous-espace à un nuage de n points, qui conduit aux calculs analytiques sans doute les plus simples est le critère classique des moindres carrés, il consiste à rendre minimale la somme des carrés des écarts :

$$\sum_i e_i c_i^2$$

L'application du théorème de Pythagore à chacun des n triangles rectangles du type $f_i o e_i$ conduit à la relation :

$$\sum_i e_i c_i^2 = \sum_i oe_i^2 - \sum_i oc_i^2$$

Donc rendre minimale $\sum_i e_i c_i^2$ est équivalent à rendre maximale $\sum_i oc_i^2$.

2.3. Composantes Principales

Considérons le système d'axes orthonormés représentant les caractères initiaux $x^1, x^2, \dots, x^P$. En projetant les individus sur une droite quelconque Δ_1 on crée un nouveau caractère C^1 dans les valeurs $(c_{11}, c_{21}, \dots, c_{n1})^t$ sont les mesures algébriques des projections des pointe e_i sur cette droite. C^1 est appelée première composante principale.

Soit le vecteur unitaire $\vec{a}$ de (Δ_1) de M- norme 1. La mesure algébrique c_1 de la projection de l'individus e_1 est alors égale au produit scolaire e_1 par a c'est à dire :

$$c_1 = {}^t e_1 M a = {}^t a M e_1 = {}^t (Ma) e_1 \text{ car } M \text{ est symétrique}$$

En posant $u = Ma$ alors on peut écrire que la composante c_1 de e_1 sur Δ vaut

$$c_1 = {}^t u e_i = \sum_{j=1}^{p} u_j x_{ij}$$

Le caractère C^1 dont les valeurs sont les n coordonnées $c_{11}, c_{21}, \dots, c_{n1}$ s'obtient alors directement par la formule $C^1 = Xu, C^1$ est donc une combinaison linéaire des p caractères initiaux au moyen du facteur u, si $M = \mathrm{I}$, il y aura égalité entre le facteur u est le vecteur unitaire a.

On note F_k le sous espace de projection. Pour cela on définit U une matrice (operateur) de projection M-orthogonal sur l'espace F_k, elle vérifie les deux conditions suivantes[10] :

1. $U^2 = U$ (U est idempotente).
2. $UM = MU^t$ (P est M-Symétrique)

Proposition 2.1. Soit $C = XU$, le tableau dont la $i^{ième}$ ligne représente la projection de l'individu e_i sur le sous espace de $\mathbb{R}^p$ engendré par les colonnes de U. Alors :

1. Le centre de gravité projeté : $g_{proj} = Pg$,
2. La matrice de covariance associée au nuage projeté : $Sproj = U^t SU$,
3. L'inertie du nuage projeté : $Iproj = trace\ (SMU^t)$.

Preuve.

1. Centre de gravité :

$$g_{proj} = C^t D 1_n$$

$$= (XU)^t D 1_n$$

$$= U^t (X^t D 1_n)$$

$$= U^t g$$

2. Matrice de covariance :

$$Sproj = C^t DC - g_{proj} g_{proj}^t$$

$$= U^t X^t DXU - U^t g g^t U$$

$$= U^t (X^t DX - g g^t) U$$

$$= U^t SU$$

3. Inertie : $I_{proj} = tr\ (S_{proj} M)$

$$= trace\ (U^t SU\ M)$$

$$= trace\ (U^t SMU^t)$$

$$= trace\ (SM(U^2)^t) = trace\ (SMU^t)$$

2.4. Axes et facteurs principaux

Nous avons vu que le 1^{er} axe principal Δ_1 avait pour propriété de rendre maximale la moyenne des carrés des distances des projections des points du nuage au centre de gravité. Ceci équivaut à rendre maximale l'inertie des projections qui vaut $p_i c_i^2$ ou les c_i sont les mesures algébriques des projections des e_i sur Δ_1 car on choisit de faire passer par le centre de gravité g du nuage. Δ_1 est l'axe d'allongement principal du nuage en ce sens que sur cet axe les c_i sont les plus dispersées possible c'est à dire : C^1 est une combinaison linéaire des x^j de variance maximale.

Pour trouver explicitement les facteurs et les axes principaux et pour alléger la démonstration on peut toujours se ramener au cas $M = I$ et en raisonnant sur le tableau des données centrées réduites transformées $Z = Y\ {}^tD_{1/s}$ avec $M = {}^tD_{1/_s}D_{1/s}$. En effet, la première composante principale de Z sera la même que celle de Y puisque les combinaisons des variables $z^1, z^2, \dots, z^P$ de variance maximale définira automatiquement la combinaison des $x^1, x^2, \dots, x^P$ de variance maximale. Si cette composante est exprimée sous la forme :

$$C^1 = Zv = Y\ {}^tD_{1/s}v = Yu,$$

Avec $u = {}^tD_{1/s}v$. Donc, la matrice de variance covariance de Z sera :

$$V_Z = {}^tYDY = R,$$

Avec R la matrice des corrélations. La composante principale C^1 à pour variance :

$$s^2_{c^1} = \mathrm{C}^{1t}DC^1 = (Zv)^tD(Zv) = (Y\ {}^tD_{1/s}u)^tD\left(Y\ {}^tD_{1/s}u\right) = v^t\ {}^tZDZv = {}^tvV_Zv = v^tRv$$

Où v est le vecteur unitaire de l'axe principal a.

Le problème consiste à trouver le vecteur v, de norme 1 tel que ${}^tvV_Zv$ soit maximal, ceci est équivalent à rendre maximal la quantité v^tV_Zv sous la contrainte $v^tv = 1$.

En utilisant la méthode des multiplicateurs de Lagrange, on pose :

$$L = v^tV_yv - \lambda(v^tv - 1)$$

Où, λ étant un multiplicateur de Lagrange, l'extrémum s'obtienne en dérivant L parraport à la variable v, nous obtenons alors :

$$2V_Zv - 2\lambda v = 0 \Rightarrow V_Zv = \lambda v$$

Le vecteur v doit être donc vecteur propre de V_Z (c'est la matrice des corrélations pour le tableau des données centré $Y = X - I_n g^t$ et c'est la matrice de variances-covariances pour le tableau des données centrées- réduites $Z = Y\ D_{1/s}$) et sa valeur propre λ doit être la plus grande puisqu'elle représente la quantité à maximiser.

La démonstration peut facilement être étendue au cas d'un deuxième vecteur unitaire v_2 orthogonal au vecteur v_1, où l'on doit annuler les dérivées du Lagrangien L, telles que :

$$L = {v_2}^t V_Z v_2 - \lambda_2({v_2}^t v_2 - 1) - \mu({v_2}^t v_1 - 0)$$

La condition d'extrémum s'écrit pour μ :

$$\frac{dL}{dv_2} = 2V_Z v_2 - 2\lambda_2 v_2 - \mu v_1 = 0$$

En multipliant les différents membres de cette relation par ${v_1}^t$, nous obtenons :

$$2{v_1}^t V_Z v_2 - 2\lambda_1 {v_1}^t v_2 - \mu {v_1}^t v_1 = 0$$

$$\Rightarrow 2\lambda_1 {v_1}^t v_2 - 2\lambda_1 {v_1}^t v_2 - \mu {v_1}^t v_1 = 0$$

$$\Rightarrow \mu = 0$$

Il reste donc comme précédemment :

$$2V_Z v_2 - 2\lambda_1 v_2 = 0$$

$$\Rightarrow V_Z v_2 = \lambda_2 v_2$$

Le vecteur v_2 doit être donc vecteur propre de V_Z et sa valeur propre λ_2 doit être la deuxième valeur la plus grande, puisqu'elle représente la quantité à maximiser.

Les axes et les facteurs principaux sont les vecteurs propres $v_1, v_2, \dots, v_p$ de la matrice des corrélations lorsque $M = I$ associer aux valeurs propres $\lambda_1, \lambda_2, \dots, \lambda_p$écrites dans un ordre décroissant. Pour trouver directement axes, facteurs et composantes principales en fonction de Y le tableaux des données centrées, il suffit d'écrire :

$$V_Z v = \lambda v \Rightarrow Z^t D Z v = (Y {D_{1/s}}^t)^t D\left(Y {D_{1/s}}^t\right) v = D_{1/s} Y^t D Y {D_{1/s}}^t v = D_{1/s} S {D_{1/s}}^t v = \lambda v$$

Et en multiplier à gauche pat ${D_{1/s}}^t$ on obtient :

$${D_{1/s}}^t D_{1/s} S {D_{1/s}}^t v = \lambda {D_{1/s}}^t$$

$$\Rightarrow MSu = \lambda u$$

L'axe a est tel que $u = Ma$ donc $MSMa = \lambda Ma$. Si M est régulière, alors on peut simplifier par M^{-1} on obtient :

$$SMa = \lambda a$$

Les axes principaux sont donc les vecteurs propres de SM et les facteurs principaux u ceux de MS,

2.5. Propriétés des axes et facteurs principaux

Comme montré précédemment, ce sont les p vecteurs propres $a_{1,}\ldots,a_p$ de la matrice SM associés à la valeurs propres $\lambda_j (j = 1 \ldots p)$, M-normé à 1, c'est-à-dire :

$$\begin{cases} SMa_j = \lambda_j a_j \\ \|a_j\|_M^2 = 1 \end{cases}$$

Proposition 2.2 : Les axes principaux satisfont les propriétés suivantes :

1. Les axes principaux a_j sont S^{-1} orthogonaux,
2. Les axes principaux a_j sont M-orthonormé.

Preuve.

1. Soit $a_j, a_{j'}$ deux axes principaux tel que :

$$< a_j, a_{j'} >_{S^{-1}} = a_j^t S^{-1} a_{j'}$$

$$= 1/\lambda_j (SMa_j)^t S^{-1} a_{j'}$$

$$= 1/\lambda_j a_j^t MSS^{-1} a_{j'}$$

$$= 1/\lambda_j a_j^t M a_{j'}$$

$$= 1/\lambda_j < a_j, a_{j'} >_M$$

$$= \begin{cases} 1/\lambda_j & si \quad j = j' \\ 0 & sinon \end{cases}$$

De même que les axes principaux, le facteur principal noté u_j est un vecteur propre de

la matrice MS associé à la valeurs propre λ_j, M^{-1} normé à 1 c'est à dire:

$$\begin{cases} MSu_j = \lambda_j u_j \\ \|u_j\|^2_{M^{-1}} = 1 \end{cases}$$

Où $u_j = Ma_j \ \epsilon \mathbb{R}^p$

Proposition 2.3 : Les facteurs principaux vérifient les propriétés suivantes :

1. u_j sont *S*-orthogonaux,

2. u_j sont M^{-1}-orthonormé,

3. u_j sont aussi les vecteurs propres de la matrice MS.

Preuve.

1. $< u_j, u_{j'} >_S = u_j^t S u_{j'}$

$$= a_j^t MSM a_{j'}$$

$$= a_j^t M \lambda_{j'} a_{j'}$$

$$= \lambda_{j'} a_j^t M a_{j'}$$

$$= \lambda_{j'} < a_j, a_{j'} >_M$$

$$= \begin{cases} \lambda_{j'} \ si \quad j = j' \\ 0 \qquad sinon \end{cases}$$

2. $< u_j, u_{j'} >_{M^{-1}} = u_j^t M^{-1} u_{j'}$

$$= a_j^t M M^{-1} M a_{j'}$$

$$= a_j^t M a_{j'}$$

$$= < a_{j'}, a_{j'} >_M$$

$$= \begin{cases} 1 \ si \quad j = j' \\ 0 \qquad sinon \end{cases}$$

3. Comme a_j est un vecteur propre de la matrice *SM*, nous avons :

$$SMa_j = \lambda_j a_j$$

$$MSMa_j = \lambda_j Ma_j$$

$$MSu_j = \lambda_j u_j$$

2.5. Propriétés des Composantes principales

Nous savons que les composants principales s'obtiennent par :

$$C = Yu$$

On observe que $MS = \mathrm{M}\, Y^t DY \Rightarrow MY^t DYu = \lambda u$. En Multipliant à gauche par Y, nous obtenons :

$$(Y\mathrm{M}Y^t D)C = \lambda C,$$

Donc C est un vecteur propre de la matrice $Y\mathrm{M}Y^t D$ associée à la même valeur propre λ.

Proposition 2.4 : La variance d'une composante principale est égale à la valeur propre λ : $V(C^i) = \lambda i$

Preuve : La variance $V(C^i)$ d'une composante principale C^i est donnée par :

$$V(C^i) = {C^i}^t D C^i = {u^i}^t Y^t DY u^i = {u^i}^t S u^i$$

Or

$$Su^i = \lambda u^i$$

Donc

$$V(C^i) = {u^i}^t S u^i = \lambda {u^i}^t u^i = \lambda$$

2.6. Interprétation des résultats d'une ACP

L'analyse en composantes principales (ACP) vise à créer de nouvelles variables, appelées composantes principales, et à les représenter graphiquement. Son objectif est de visualiser les relations entre ces composantes principales, ainsi que de détecter d'éventuels regroupements d'individus et variables.

2.6.1. Interprétation des individus

A. Qualité de représentation sur l'espace des projections

Cette mesure représente l'aplatissement du nuage sur le plan principal, elle calculée par un quotient représentant l'inertie cumulé des k premiers axes principaux :

$$QLT = \frac{\lambda_1+\lambda_2+\ldots.+\lambda_k}{\lambda_1+\lambda_2+\ldots.+\lambda_P} = \frac{\lambda_1+\lambda_2+\ldots.\lambda_k}{I_g}, k \leq p.$$

Plus cette qualité se rapproche de 1, meilleure est la qualité de la représentation dans l'espace principal.

B. Qualité de représentation d³un individu e_i par rapport à l'axe (Δ_k)

On mesure la qualité de la projection d'un individu e_i sur Δ_k par le carré du cosinus de l'angle θ_i formé entre le vecteur ge_i et l'axe Δ_k :

$$QLT\text{l}\,(e_i) = cos^2(\theta_i) = \frac{(\overrightarrow{oc_l})^2}{(\overrightarrow{oe_l})^2}$$

Donc,

$$cos^2(\Theta_i) = \frac{\boldsymbol{c_{il}^2}}{\|\boldsymbol{z_i}\|^2}$$

En général, on mesure la qualité de la projection d'un individu e_i sur deux axes l et l' par le carré du cosinus de l'angle $\Theta_{i(l\,;l')}$ entre le vecteur z_i et sa projection orthogonale sur (l,l') :

$$QLT_{l,l'}(e_i) = cos^2(\Theta_{i(l,l')}),$$

Avec

$$cos^2(\Theta_{il}) = \frac{\boldsymbol{c_{il}^2+c_{il'}^2}}{\|\boldsymbol{z_i}\|^2}$$

On peut donc dire que : $QLT_{l,l'}(i) = QLT_l(i) + QLT_{l'}(i)$.

Plus la valeur du $\cos^2$ est proche de 1, plus la représentation graphique de l'individu est demeilleure qualité.

2.6.2. Interprétation des variables (Cercle des corrélations)

Pour donner une signification à la composante principale C^l, il faut la relier aux variables initiales x^j , en calculant le coefficient de corrélation $r(x^j, C^l)$ et on s'intéresse au plus fort coefficient en valeur absolue. Chaque variable représentée par les cordonnées : $(r(C^1, x^j), r(C^2 2, x^j))$ est dans un cercle de corrélation de rayon 1. On exprime la qualité de représentation d'une variable quantitative x^j sur le $l^{iéme}$ axe factoriel par le coefficient de corrélation linéaire $r(c^l, x^j)$, entre la variable initiale x^j et la composante principale C^l.

Proposition 2.5 : Le coefficient de corrélation entre une variable x^j et une composante principale C^l est : $r(C^l, x^j), = \sqrt{\lambda_l} u_{jl}$.

Preuve.

Comme $r(C^l, x^j), = r(C^l, z^j), = \frac{\boldsymbol{cov}(C^l, z^j)}{\boldsymbol{S_{C^l} S_{zj}}}$, Alors :

$$cov(z^j \ , C^l) \ = \ z^{j^t} DC^l = \left(z^j\right)^t DZu^l$$

$$= \lambda_l u_{jl}, \text{ car } Z^t DZ = R \text{ et } Ru_l = \lambda_l u_l.$$

Donc,

$$r(C^l, z^j) = \frac{\lambda_l u_{jl}}{S_{C^l} S_{zj}}$$

$$= \frac{\lambda_l u_{jl}}{\sqrt{\lambda_l}}$$

$$= \sqrt{\lambda_l} u_{jl}.$$

2.7. Conclusion

Dans ce chapitre, nous avons présenté l'Analyse en Composantes Principales (ACP) comme une méthode fondamentale en statistique exploratoire multidimensionnelle. L'objectif de cette méthode est d'obtenir une représentation simplifiée du nuage de données qui se rapproche davantage de la réalité dans un espace de dimension réduite. Cela permet

d'étudier la similarité entre les individus et la corrélation entre les variables, où ces informations pertinentes sont résumées et visualisées dans un tableau de données.

CHAPITRE III
Analyse Factorielle des Correspondances

Dans ce chapitre, nous aborderons l'Analyse Factorielle des Correspondances (AFC), une méthode statistique largement utilisée pour identifier les relations entre différentes variables, notamment lorsqu'elles sont catégorielles. L'objectif de l'AFC est de réduire la complexité des données en explorant les relations entre les variables multidimensionnelles.

L'AFC offre une approche puissante pour analyser des ensembles de données complexes et pour visualiser les relations entre les variables d'intérêt. En comprenant les étapes clés de l'analyse et en interprétant correctement les résultats, nous pourrons tirer des conclusions significatives et prendre des décisions éclairées.

3.1. Présentation de la méthode

L'Analyse Factorielle des Correspondances (AFC) s'applique aux tableaux de contingence ou tableaux croisés. Un tel tableau est une représentation des effectifs des individus en fonction de deux caractères qualitatifs. Le tableau que l'on désigne par N a une taille de p lignes et q colonnes, où p représente le nombre de modalités pour le premier caractère et q représente le nombre de modalités pour le deuxième caractère. Chaque cellule du tableau contient le nombre n_{ij}, qui représente le nombre d'individus ayant à la fois la modalité i du premier caractère et la modalité j du deuxième caractère. Ces nombres d'individus constituent les effectifs observés dans chaque cellule du tableau croisé.

Prenons l'exemple d'un disquaire qui répartit la vente de 1000 disques, suivant la catégorie de la musique (trois modalités: chansons(C), Jazz(J) et musique classique(Mc)) et la population des utilisateurs (quatre modalités: jeunes sans distinction du sexe(JSDS),

adultes féminins(AF), adultes masculins(Ms) et personnes âgées sans distinction de sexe(PASDS)), il a obtenu le tableau suivant [12]:

Tableau 3.1 : Tableau de contingence pour l'exemple d'un disquaire

	JSDS	**AM**	**AF**	**PASDS**	**Total**
C	69	172	133	27	401
Ja	41	84	118	11	254
MC	18	127	157	42	345
Total	128	383	408	81	1000

L'objectif de l'Analyse Factorielle des Correspondances (AFC) appliquée à un tableau de contingence est d'analyser la structure de dépendance entre les deux caractères qualitatifs et de mettre en évidence les principales tendances de cette dépendance. Le tableau de contingence peut être lu selon ses lignes ou ses colonnes, ce qui permet de mettre en évidence des aspects différents de la dépendance entre les variables. La lecture selon les lignes met en évidence les associations entre les modalités du premier caractère et du deuxième caractère, tandis que la lecture selon les colonnes met en évidence les associations entre les modalités du deuxième caractère et du premier caractère. Ces deux lectures permettent d'obtenir une compréhension plus complète de la structure de dépendance dans le tableau de contingence.

a. Si on désire savoir pour chaque catégorie de musique comment se répartit la population des utilisateurs, on calculera les pourcentages en lignes en le divisant les effectifs n_{ij} de la ligne $n°i$ par le total $n_{i.}$ de la ligne. On obtient se qu'on appelle tableau de profile ligne :

Tableau 3.2 : Tableau des profils des lignes.

	ASDS	**AF**	**AM**	**PA**	$\frac{n_{i\cdot}}{n}$
C	0.17207	0.42893	0.33167	0.06733	0.401
Ja	0.16142	0.33071	0.46457	0.04331	0.254
MC	0.05217	0.36812	0.45507	0.12464	0.345
$n_{\cdot j}/n$	0.128	0	0.408	0.081	1

Le profil marginal $n_{\cdot j}$ est aussi le profil moyen car il est la moyenne des lignes pondérés par le poids $p_i = \frac{n_{i\cdot}}{n}$ de chaque ligne.

$$\bar{x}^j = \sum_{j=1}^{q} p_i \frac{n_{ij}}{n_{i\cdot}} = \sum_{j=1}^{q} \frac{n_{i\cdot}}{n} \frac{n_{ij}}{n_{i\cdot}} = \frac{1}{n} \sum_{j=1}^{q} n_{ij} = \frac{n_{\cdot j}}{n}$$

b. Si réciproquement ont veut savoir pour une catégorie de population données, comment se répartissent les différentes catégories de musique, on calculera les profile des colonnes en divisant les effectifs n_{ij} de la colonne j par $n_{\cdot j}$ le total de la colonne j. Le tableau des profils colonnes est donné dans le tableau suivant :

Tableau 3.3 : Tableau des profils des colonnes.

	JSDS	**AM**	**AF**	**PASDS**	$\frac{n_{i\cdot}}{n}$
C	0,539	0,449	0,326	0,333	0,401
Ja	0,320	0,219	0,289	0,136	0,254
MC	0,141	0,332	0,385	0,516	0,345
$n_{\cdot j}/n$	0,128	0,383	0,408	0,81	1

Si on appelle D_1 et D_2 les matrices diagonales des effectifs marginaux :

$$D_1 = \begin{bmatrix} n_{1\cdot} & & 0 \\ & \ddots & \\ 0 & & n_{p\cdot} \end{bmatrix} \quad ; \quad D_2 = \begin{bmatrix} n_{\cdot 1} & & 0 \\ & \ddots & \\ 0 & & n_{\cdot q} \end{bmatrix}$$

Le tableau renfermant les p profils des lignes est le produit matriciel :

$$D_1^{-1}N = \left(\frac{n_{ij}}{n_{i\cdot}}\right) \; i = 1 \ldots p \text{ et } j = 1 \ldots q$$

Le tableau des profils des colonnes est le produit matriciel :

$$ND_2^{-1} = \left(\frac{n_{ij}}{n_{\cdot j}}\right) \; i = 1 \ldots p \text{ et } j = 1 \ldots q$$

3.1.1. Analyse dans $\mathbb{R}^q$

Si on s'intéresse aux lignes de tableau N, on peut considérer le tableau $D_1^{-1}N$ des profils des lignes comme un tableau individus caractères particulier, et effectuer une analyse en composantes principales. Les individus de cette analyse sont les profils des lignes, chaque individu i ayant pour coordonnées les quantités $\frac{n_{ij}}{n_{i\cdot}}$ (pour $j = 1 \ldots q$), affecté de la masse $f_{i\cdot} = \frac{n_{i\cdot}}{n}$. L'ACP revient alors à étudier la dispersion du nuage des p profils dans R^q autour de leur centre de gravité qui n'autre que le profil marginal colonne $\left(\frac{n.1}{n}, \frac{n.2}{n}, \ldots, \frac{n.q}{n}\right)$. En d'autres termes on cherche a rendre compte de l'écartement entre les $\frac{nij}{ni\cdot}$ et les $\frac{n\cdot j}{n}$, ce qui est une façon d'analyser la dépendance entre les deux caractères qualitatifs [13].

3.1.2. Analyse dans $\mathbb{R}^p$

Inversement si on s'intéresse aux colonnes de N, c'est à dire le tableau ND_2^{-1} ou plutôt son transposé $D_2^{-1}N^T$ qui jouera de rôle de tableau individus-caractères, on étudiera alors la configuration des profils des colonnes dans R^P. Chaque individu, j ayant pour coordonnées les quantités $\frac{n_{ij}}{n_{\cdot j}}$ (pour $i = 1,\ldots, p$), affecté de la masse $f_j = \frac{n_{i\cdot}}{n}$.

3.2. Analyse en composantes principales des tableaux des deux profils

Pour effectuer une ACP sur ces tableaux, vous devez définir une formule de distance entre les objets, en d'autres termes, une métrique**.**

3.2.1. La métrique χ^2(khi − deux)

On appelle métrique de χ^2 pour les lignes, la matrice diagonale :

$$M_1 = nD_2^{-1} = \begin{pmatrix} \frac{n}{n_{\cdot 1}} & & 0 \\ & \ddots & \\ 0 & & \frac{n}{n_{\cdot q}} \end{pmatrix}_{q \times q}$$

Elle est définie par l'inverse du profil marginal des colonnes de N. De la même façon, la métrique de χ^2 pour les colonnes, est définie par l'inverse du profil marginal des lignes de N.

$$nD_1^{-1} = \begin{pmatrix} \frac{n}{n_{1\cdot}} & & 0 \\ & \ddots & \\ 0 & & \frac{n}{n_{p\cdot}} \end{pmatrix}_{p \times p}$$

3.2.2. ACP du nuage des profils lignes

Afin de se ramener à la métrique usuelle, on peut écrire la formule exprimant la distance entre profils lignes de manière suivante :

La formule exprimant la distance entre deux profils lignes i et k s'écrit :

$$d^2 = \sum_{j=1}^{q} \frac{n}{n_{\cdot j}} \left(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{kj}}{n_{k\cdot}} \right)^2$$

$$= \sum_{j=1}^{q} \frac{1}{\frac{n_{\cdot j}}{n}} \left(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{kj}}{n_{k\cdot}} \right)^2$$

$$= \sum_{j=1}^{q} \left[\frac{1}{\sqrt{\frac{n \cdot j}{n}}} \left(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{jj}}{n_{k\cdot}} \right) \right]^2$$

Si nous posons : $f_{ij} = \frac{n_{ij}}{n}, f_i = \frac{n_{i\cdot}}{n}$ et $f_j = \frac{n_{\cdot j}}{n}$ alors on obtient :

$$d^2(i,k) = \sum_{j=1}^{q} \left[\left(\frac{f_{ij}}{\sqrt{f_{\cdot j}} f_{i\cdot}} - \frac{f_{ij}}{\sqrt{f_{\cdot j}} f_{k\cdot}} \right) \right]^2$$

Cette dernière formule permet de se ramener à la métrique usuelle où le profil ligne devient alors : $\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}$. Calculons les coordonnées du centre de gravité du nuage :

$$g_j = \sum_{i=1}^{p} p_i \frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}$$

$$= \sum_{i=1}^{p} f_{i\cdot} \frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}$$

$$= \sqrt{f_{\cdot j}}$$

Les coordonnées centrées sont :

$$\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right)$$

Appliquons au tableau des profils lignes X de terme général $x_{ij} = \left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right)$ les résultats du deuxième chapitre sur l'ACP. Les facteurs principaux sont les vecteurs propres de , la metrique est ici ND_2^{-1}, et la matrice V est égale au produit matriciel $X^t DX$ où (X est tableau du profils ainsi obtenus et D la matrice diagonale des poids $d_{ii} = \frac{n_{i\cdot}}{n} = f_{i\cdot}$).

Donc nous avons : $T = MV = X^t DX$

$$t_{ik} = \sum_{j=1}^{q} f_{i\cdot} \left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right) \left(\frac{f_{kj}}{f_{k\cdot}\sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right)$$

Il est possible de donner à cette matrice T une forme simple posons en effet :

$$x_{ij}^* = \left(\frac{f_{ij} - f_{i\cdot} f_{\cdot j}}{\sqrt{f_{i\cdot} f_{\cdot j}}} \right)$$

Alors la matrice T à diagonaliser s'exprime en fonction du tableau noté X^*. $T = X^{*t}X^*$. Nous projetons le point i sur l'axe factorielle U_a en obtient les coordonnée de profils lignes.

$$C^{\alpha} = \sum_{i=1}^{f_1} \left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right) U_{\alpha}^{j}$$

3.3. ACP du nuage de points dans $\mathbb{R}^p$

Les ensembles mis en correspondance dans le tableau des fréquences jouent des rôles analogues. L'analyse dans $\mathbb{R}^p$ peut donc se déduire de celle menée dans $\mathbb{R}^q$ par permutation des rôles des indices i et j. Ainsi les coordonnées du point j seront maintenant les quantités $\frac{fij}{fi\cdot}\sqrt{f_{i\cdot}}$ ce point j sera muni de la masse $f_{\cdot j}$. La $i^{\text{ème}}$ coordonnée du centre de gravité H du nuage des q points s'écrit: $h_i = \sqrt{f_{i\cdot}}$.

La matrice MV a diagonaliser aura de terme général :

$$W_{ik} = \sum_{j=1}^{q} f_{\cdot j} \left(\frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}} - \sqrt{f_{i\cdot}} \right) \left(\frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}} - \sqrt{f_{i\cdot}} \right)$$

Donc la matrice MV est le produit matriciel de $W = X^{*t}X^*$. En fin les coordonnées des profils lignes s'obtiennent par la formule suivante :

$$C^{\alpha} = \sum \left(\frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}} - \sqrt{f_{i\cdot}} \right) V_{\alpha}^{i}$$

3.4. Relation entre les deux espaces $\mathbb{R}^q$ et $\mathbb{R}^p$

On montre qu'il existe une dualité entre les deux analyses et que les matrices T et W partagent les mêmes valeurs propres. Il est ainsi observé que l'Analyse en Composantes Principales (ACP) appliquée au nuage de profils des lignes est équivalente à l'ACP appliquée au nuage de profils des colonnes. Les facteurs principaux obtenus dans une analyse sont, à une racine carrée près de $\sqrt{\lambda}$, les composantes principales de l'autre [13].

$$U_\alpha = \frac{1}{\sqrt{\lambda_a}} X^* V_\alpha$$
$$V_\alpha = \frac{1}{\sqrt{\lambda_a}} X^* U_\alpha$$

3.5. La représentation graphique

3.5.1. Optique ACP

Si on considère les profils lignes comme individus ($1^{\text{ière}}$ ACP) il est naturel de représenter les modalités du premier caractère par les cordonnées de ces profils sur les axes principaux. Or les composantes principales s'obtiennent en multipliant les vecteurs propres Ua par leurs valeurs propres associées $\sqrt{\lambda_\alpha}$.

Inversement, la deuxième ACP sur les profils des colonnes conduit à représenter les modalités du deuxième caractère qualitatif, en multipliant les vecteurs propres V_a par leurs valeurs propres associées $\sqrt{\lambda_\alpha}$, on obtient alors deux représentations séparées des modalités de chaque caractère.

3.5.2. La représentation Simultanées Usuelle

Elle consiste à représenter les modalités du premier caractère $i\ (1 \cdots p)$ par les points de cordonnées $\sqrt{\lambda_\alpha} U_\alpha^{\text{i}}$ et les modalités j $(1 \cdots q)$ du deuxième caractère par les points de cordonnées $\sqrt{\lambda_a} V_a^{\text{j}}$. Ceci revient à superposer les graphiques des deux ACP, opération dont la justification mathématique est délicate dans le cadre de l'ACP puisqu'on mélange sur un même graphique des individus et des caractères, éléments d'ensembles différents.

3.5.3. L'étude des contributions

Pour interpréter correctement les graphiques, il faut comme en ACP tenir compte, d'une part, de la proximité entre les points et les plans principaux et d'autre part, du rôle joué par chaque point dans la détermination d'un axe. Les données étant qualitatives on n'utilise ici la corrélation entre caractères et axes principaux, les contributions seront :

- **Contribution des points à l'Inertie des axes**

Les cordonnées des modalités sur les axes étant $\sqrt{\lambda_\alpha}U_a^{\mathrm{i}}$ et $\sqrt{\lambda_\alpha}V_\alpha^j$, l'inertie au $\lambda_a k^{\text{ième}}$ axe peut se décomposer selon les modalités du premier caractère ou celle du second :

$$\lambda_\alpha = \sum_{i=1}^{p} f_{i\cdot}\left(\sqrt{\lambda_\alpha}U_a^i\right)^2 = \sum_{j=1}^{q} f_{\cdot j}\left(\sqrt{\lambda_\alpha}V_\alpha^j\right)^2$$

La part de due à la modalités i est donc $f_{i\cdot}\left(\mathrm{u}_a{}^{\mathrm{i}}\right)^2$, c'est la contribution de la modalité i à l'axe α (souvent appelée improprement contribution absolue), pour interpréter les axes, on cherche les contributions les plus importantes.

- **Proximité entre les points et les axes principaux**

Comme en ACP on utilise le cosinus carré de l'angle entre les individus, ici les profils ligne et les profils colonnes et l'axe principale pour mesurer la qualité de la représentation dans les plans principaux. La somme de ces cosinus carré pour un même individu sur tous les axes égaux à 1.

3.6. Conclusion

Ce chapitre a permis de détailler la méthode de l'Analyse Factorielle des Correspondances (AFC). L'AFC s'est révélée être une technique puissante pour analyser les relations entre les variables catégorielles et identifier les tendances et les associations dans les données. De plus, l'AFC offre des outils statistiques pour évaluer la signification des résultats et interpréter les contributions des variables aux dimensions principales. En conclusion, l'Analyse Factorielle des Correspondances est un outil essentiel pour explorer et comprendre les données catégorielles, offrant de nouvelles perspectives pour la recherche et l'analyse dans divers domaines d'application.

CHAPITRE IV

L’utilisation de l’ACP pour la Reconstitution des Logs des Puits du Champ de MESDAR au Sein de l’Entreprise SONATRACH

Dans ce chapitre, nous avons présenté notre expérience pratique lors d'un stage d'un mois au sein de la Division Petroleum Engineering & Development de la société nationale SONATRACH. L'objectif principal de ce stage était de mettre en pratique nos connaissances et d'acquérir une meilleure compréhension des réalités de l'entreprise.

Dans l'industrie pétrolière et gazière, l’ACP peut être utilisée pour analyser des ensembles de données complexes de forage, de production et d'ingénierie de réservoir afin de mieux comprendre et améliorer la prise de décision.

Un exemple de la façon dont l’ACP peut être utilisée dans l'industrie pétrolière et gazière est d'analyser les données de diagraphie de puits, qui fournissent des informations sur les propriétés souterraines d'un réservoir. Les diagraphies de puits enregistrent des mesures de diverses propriétés physiques des formations rocheuses, telles que la résistivité, la porosité et la densité. Ces propriétés peuvent être utilisées pour estimer la quantité d'huile ou de gaz dans le réservoir, ainsi que le potentiel de production d'un puits

Cependant, les données de diagraphie de puits peuvent être très volumineuses et complexes, avec de nombreuses variables qui sont fortement corrélées. L’ACP peut être utilisée pour identifier les variables les plus importantes et réduire la dimensionnalité de l'ensemble de données, ce qui facilite son analyse et sa visualisation. En réduisant la dimensionnalité de l'ensemble de données, l’ACP peut également aider à identifier les relations entre les variables qui peuvent ne pas être apparentes dans les données originales.

4.1. Contexte et problématique

Notre travail avait pour but de traiter les problèmes de qualité des interprétations liés à la digitalisation des logs des anciens puits du champ de MESDAR, ainsi qu'à l'état du trou. Un décalage entre la porosité carotte (CPOR) et la porosité effective (PHIE) a été observé dans la partie supérieure du réservoir Cambrien de certains puits, à savoir MDR-10, MDR-13, MDR-20, MDR-23 et MDR-7.

Deux solutions ont été envisagées pour résoudre ce problème. La première solution consistait à faire confiance aux données carotte au détriment des logs, en appliquant un lissage des données carotte (Porosity Windowing). Cependant, la deuxième solution, qui a été retenue, était de reconstituer les logs affectés en utilisant l'Analyse en Composante Principales (ACP).

4.2. Méthodologie et Analyse en ACP

Nous rappelons que l'ACP est une méthode permettant de transformer des variables corrélées en nouvelles variables indépendantes appelées "composantes principales". Elle permet de réduire le nombre de variables et de rendre l'information moins redondante. Dans le cadre de ce stage, la méthode ACP a été appliquée avec succès sur l'ensemble des puits mentionnés précédemment.

Le workflow utilisé pour l'analyse en ACP se résume comme suit (Figure 4.1) :

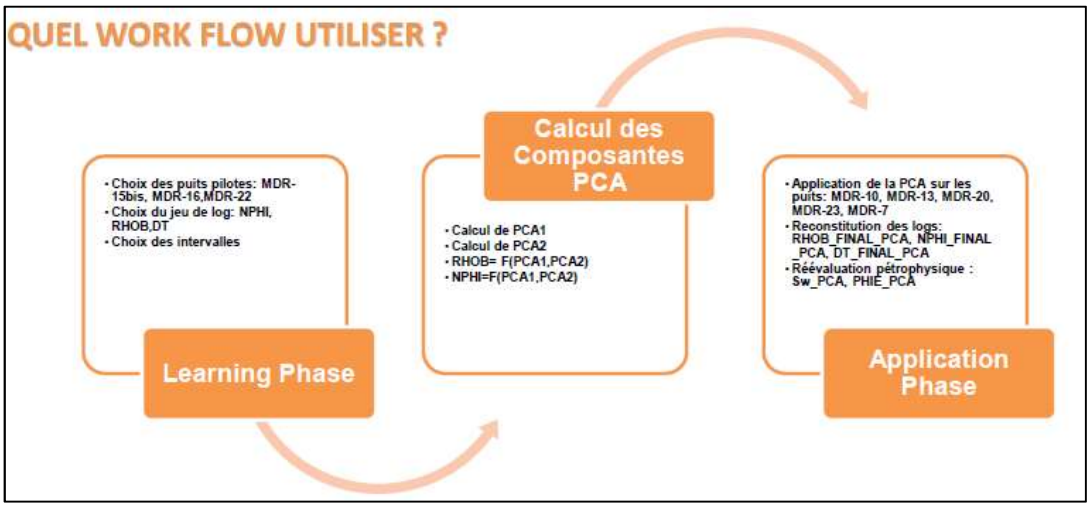

Figure 4.1 : Workflow utilisé pour l'analyse en ACP

4.2.1. Phase d'apprentissage

Dans cette phase, nous avons sélectionné des puits pilotes pour la reconstitution des logs. Les puits choisis étaient MDR-16, MDR-22 et MDR-15b, en raison de la qualité des données de diagraphies et de la bonne corrélation entre la porosité effective (PHIE) et la porosité carotte (CPOR). Un jeu de logs spécifique (NPHI, RHOB et DT) a été sélectionné pour la reconstitution, en veillant à éliminer les intervalles de mauvaise qualité par le biais d'un contrôle qualité (QC).

4.2.2. Calcul des composantes principales d'une ACP

La méthode ACP a été appliquée en utilisant une application dénommée interactive Petrophysics (IP). Les logs ont été reconstitués en utilisant les calculs des composantes principales PCA1, PCA2 et PCA3 : RHOB = F(PCA1, PCA2, PCA3), NPHI = F(PCA1, PCA2,PCA3) et DT = F(PCA1, PCA2,PCA3). Les figures 4.2 et 4.3 illustrent le processus de calcul des composantes principales PCA1, PCA2 et PCA3 du log reconstitué RHOB_PCA du puits MDR-16.

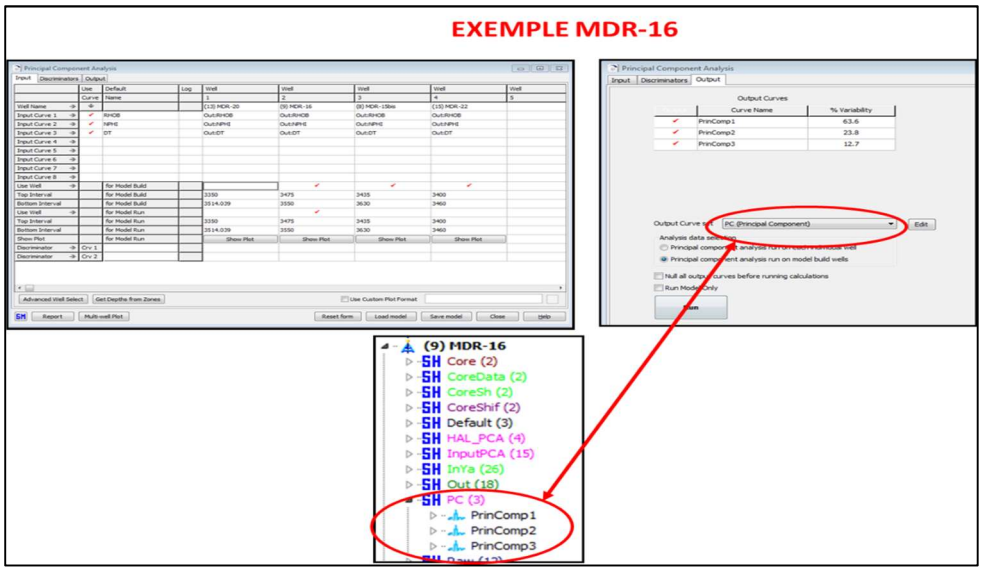

Figure 4.2 : Composantes principales PCA1, PCA2 et PCA3 du log reconstitué RHOB_PCA du puits MDR-16

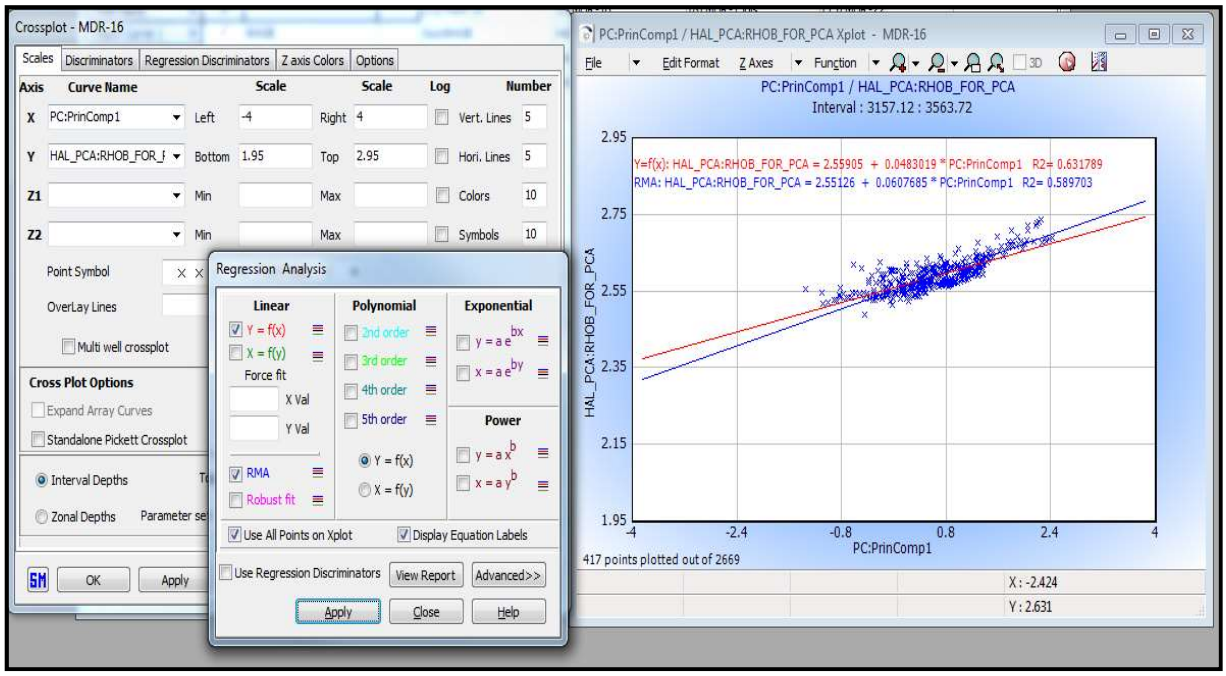

Figure 4.3 : Composante principale PCA1 du log reconstitué RHOB_PCA du puits MDR-16

4.2.3. Phase d'application

La méthode ACP a été appliquée aux puits MDR-10, MDR-13, MDR-20, MDR-23 et MDR-7. Les logs reconstitués du puits MDR-20 sont illustrés dans la figure 4.4, notamment les courbes RHOB_FINAL_PCA, NPHI_FINAL_PCA et DT_FINAL_PCA. Une bonne corrélation a été observée entre les logs prédits et les logs bruts.

Une réévaluation pétrophysique a été réalisée en utilisant les courbes reconstituées par ACP. Les résultats du puits MDR-20 ont été présentés dans la figure 4.5, incluant la saturation SW_PCA et la porosité effective PHIE_PCA. Pour plus de commodité, le lecteur trouvera l'ensemble des résultats relatifs à l'application de la méthode ACP des autres puits en annexe 1.

Les nouvelles interprétations n'ont démontré aucun changement significatif, ce qui permet de lever l'incertitude électrique liée aux logs. Toutefois, les carottes demeurent suspectes dans la partie supérieure du réservoir.

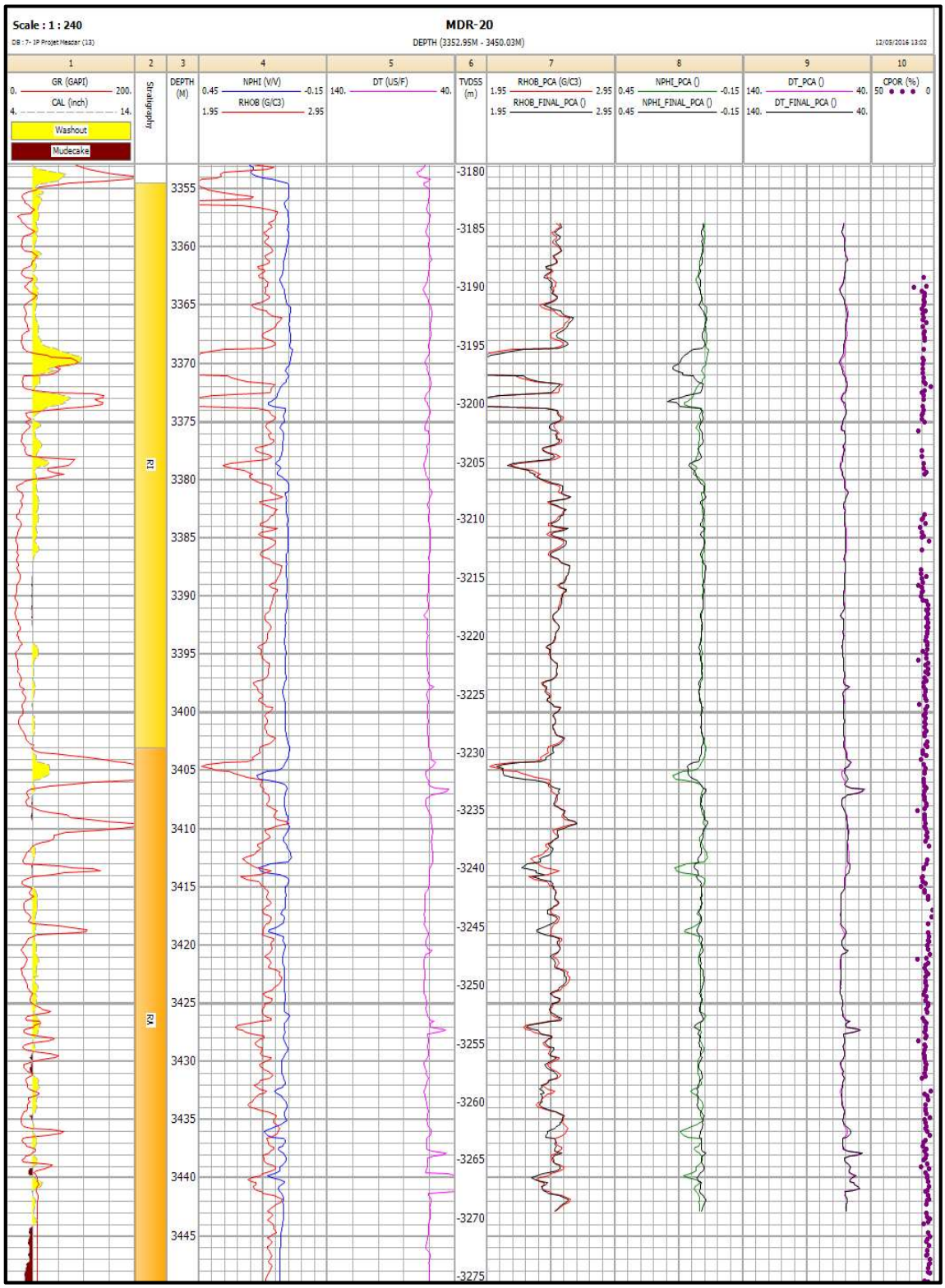

Figure 4.4 : Well composite du puits MDR-20.

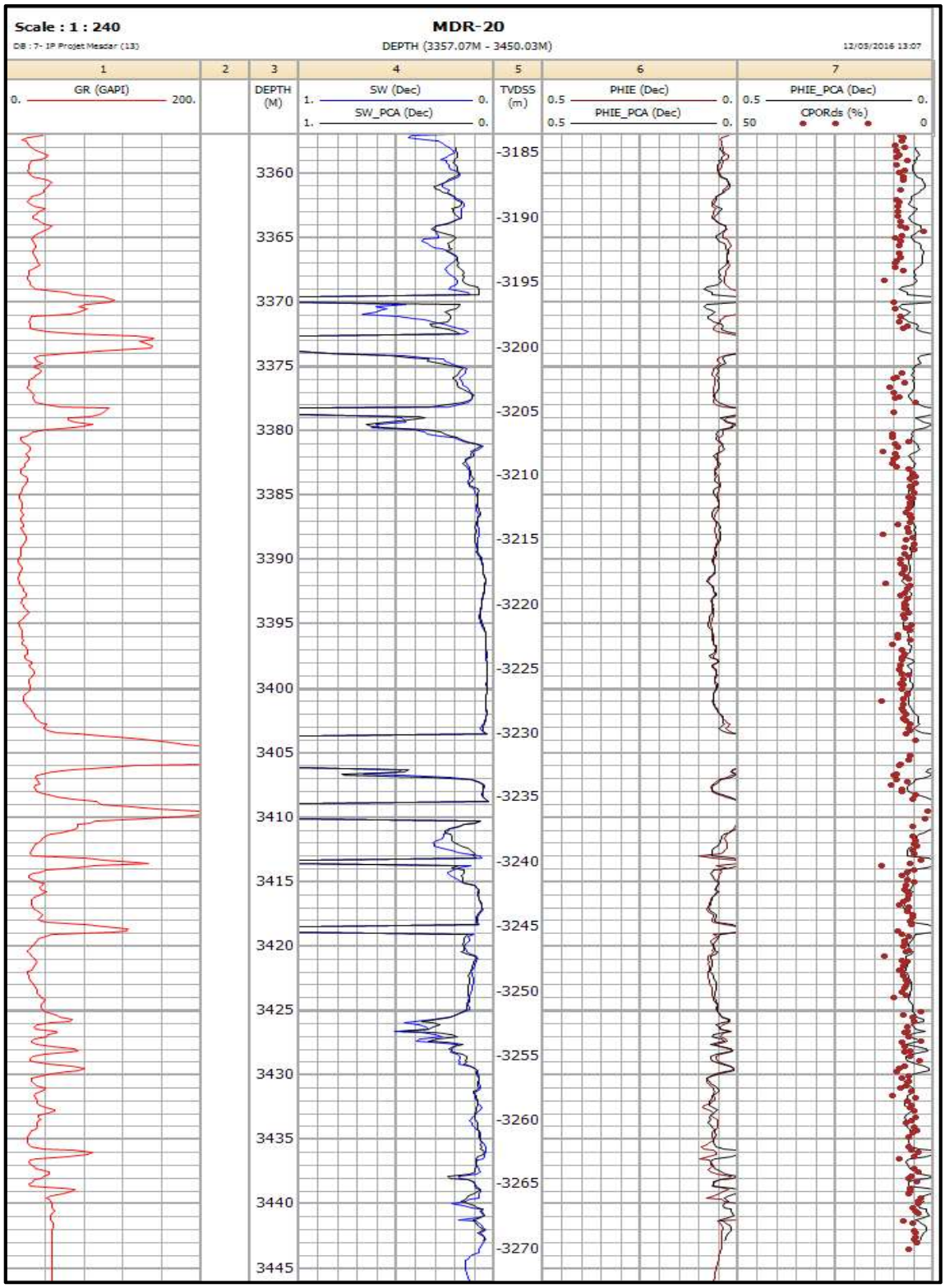

Figure 4.5 : Réévaluation pétro-physique du puits MDR-20 en utilisant les courbes reconstituées par ACP

4.3. Conclusion

En résumé, l'ACP s'est révélée être une technique polyvalente et efficace pour la reconstitution des logs dans l'industrie pétrolière et gazière. En identifiant les variables les plus importantes et en réduisant la dimensionnalité des données, l'ACP permet aux professionnels de mieux comprendre les caractéristiques des réservoirs et de prendre des décisions plus éclairées.

Nous avons également pu observer les avantages de l'application de l'ACP dans le domaine de l'interprétation des logs. Cette méthode permet non seulement de reconstituer les données manquantes ou corrompues, mais elle offre également la possibilité d'identifier les relations entre les variables et de détecter des tendances significatives.

Cependant, il convient de noter que malgré les résultats positifs obtenus grâce à l'analyse en ACP, les carottes dans la partie supérieure du réservoir demeurent suspectes. Cela souligne l'importance de prendre en compte différents facteurs et de combiner plusieurs méthodes d'analyse pour obtenir une évaluation complète et précise des caractéristiques du réservoir.

CHAPITRE V

Description de l'Application NUMIDATA

L'application de l'analyse des données est facilitée par l'utilisation de logiciels spécifiques. L'interprétation de l'analyse des données nécessite de nombreux calculs et de nombreux graphiques. Là encore, l'application NumiData que nous avons réalisé facilite l'utilisation de l'analyse en composantes principales et l'analyse factoriel des correspondances. Elle permet d'effectuer les calculs nécessaires à l'interprétation des résultats et elle donne la possibilité de tracer de nombreux graphiques qui illustrent les résultats. Il est ainsi possible de se rendre compte des possibilités offertes par un logiciel d'Analyse des données. Le lecteur intéressé pourra l'utiliser, en choisissant l'une des méthodes programmées pour différentes études de domaines très variés

Il existe plusieurs logiciels d'Analyse des Données qui ont chacun leurs avantages et leurs inconvénients. La version de ce logiciel donne une bonne idée de l'apport de l'informatique à l'Analyse des données et de l'aide qu'un utilisateur peut en attendre.

Ce chapitre présentera en détail les fonctionnalités essentielles de l'application NumiData, notamment son interface utilisateur et les résultats obtenus lors de l'analyse des données.

5.1. Interface Utilisateur

L'interface utilisateur de NumiData est soigneusement conçue pour offrir une expérience fluide et conviviale. Elle est développée en utilisant les puissants frameworks Tkinter et PyQt5, garantissant une navigation intuitive et une interaction agréable avec l'application. Voici une présentation des éléments clés de l'interface, conçus pour faciliter la manipulation des données et l'analyse des résultats. L'Application étant installée sur le disque dur, dans le répertoire choisi. Double clic sur le répertoire de l'Application, On obtient la fenêtre d'ouverture suivante (Figure 5.1) :

Figure 5.1 : Fenêtre d'Accueil

5.2. Barre de tâches

Dans la barre de tâches, vous trouverez trois boutons principaux qui permettent d'accéder rapidement aux fonctionnalités essentielles de NumiData.

5.2.1. DATA :

Ce bouton permet aux utilisateurs de choisir entre deux options : "Upload" ou "Remplir manuellement". L'option "Upload" permet d'importer des données à partir de fichiers externes, tandis que l'option "Remplir manuellement" permet aux utilisateurs d'entrer les données directement dans l'application.

Figure 5.2 : Menu Data.

5.2.2. Method

Ce bouton permet aux utilisateurs de sélectionner la méthode d'analyse de leur choix : "ACP" (Analyse en Composantes Principales) ou "AFC" (Analyse Factorielle des Correspondances). Sous l'option "ACP", les utilisateurs ont la possibilité de choisir entre les variantes "Normé" ou "Non normé".

Figure 5.3 : Menu Method.

5.2.2.1. Analyse en Composantes Principales (ACP) :

L'ACP est une technique d'analyse statistique qui permet de réduire la dimensionnalité des données tout en préservant au maximum leur structure. NumiData offre deux options d'ACP : "Normé" et "Non normé". Voici les résultats obtenus pour chaque option :

a. ACP Normé :

Lorsque l'option "ACP Normé" est sélectionnée, les résultats suivants sont calculés et affichés :

- Matrice de corrélation des variables : Cette matrice indique la corrélation entre les différentes variables,
- Matrice centrée : Cette matrice représente les données centrées autour de leur moyenne,
- Matrice centrée réduite : Cette matrice est obtenue en divisant la matrice centrée par l'écart type des variables,
- Valeurs propres et pourcentage de contribution des deux premières valeurs propres. Ces valeurs indiquent l'importance relative des composantes principales,
- Vecteurs propres correspondant aux valeurs propres : Ces vecteurs indiquent la direction des composantes principales,
- Cosinus carrée de thêta : Cet indice mesure la qualité de représentation des variables dans le plan factoriel,
- Coordonnées des variables : Ces coordonnées montrent la contribution des variables à chaque composante principale,
- Projection des coordonnées des variables dans le cercle de corrélation : Ce graphe visualise la relation entre les variables dans le plan factoriel (Figure 5.5).

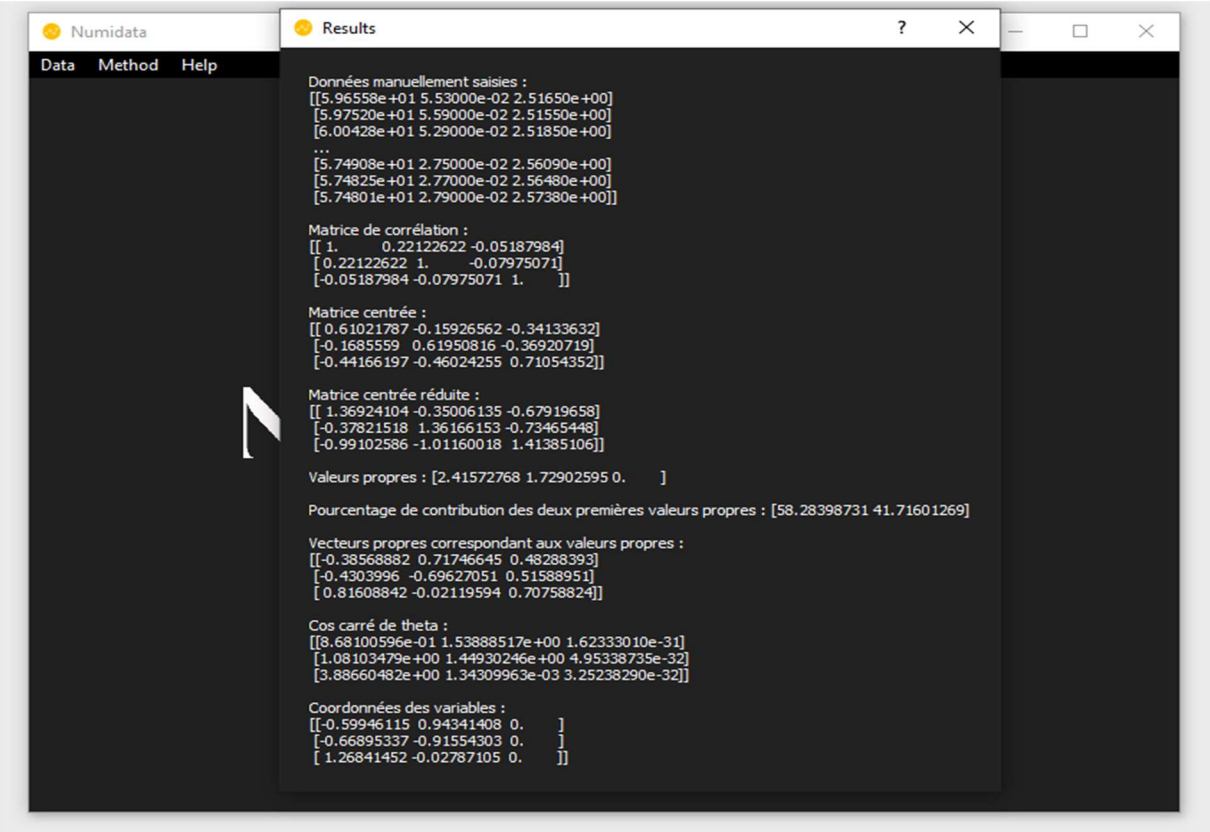

Figure 5.4 : Résultats d'une ACP normée.

b. ACP Non normé

L'option "ACP Non normé" fournit des résultats similaires à l'ACP normé, à l'exception de l'utilisation de la matrice centrée réduite. Les résultats incluent :

- Matrice de corrélation des variables,
- Matrice centrée,
- Valeurs propres et pourcentage de contribution des deux premières valeurs propres,
- Vecteurs propres correspondant aux valeurs propres,
- Coordonnées des variables,
- Projection des coordonnées des variables dans le cercle de corrélation.

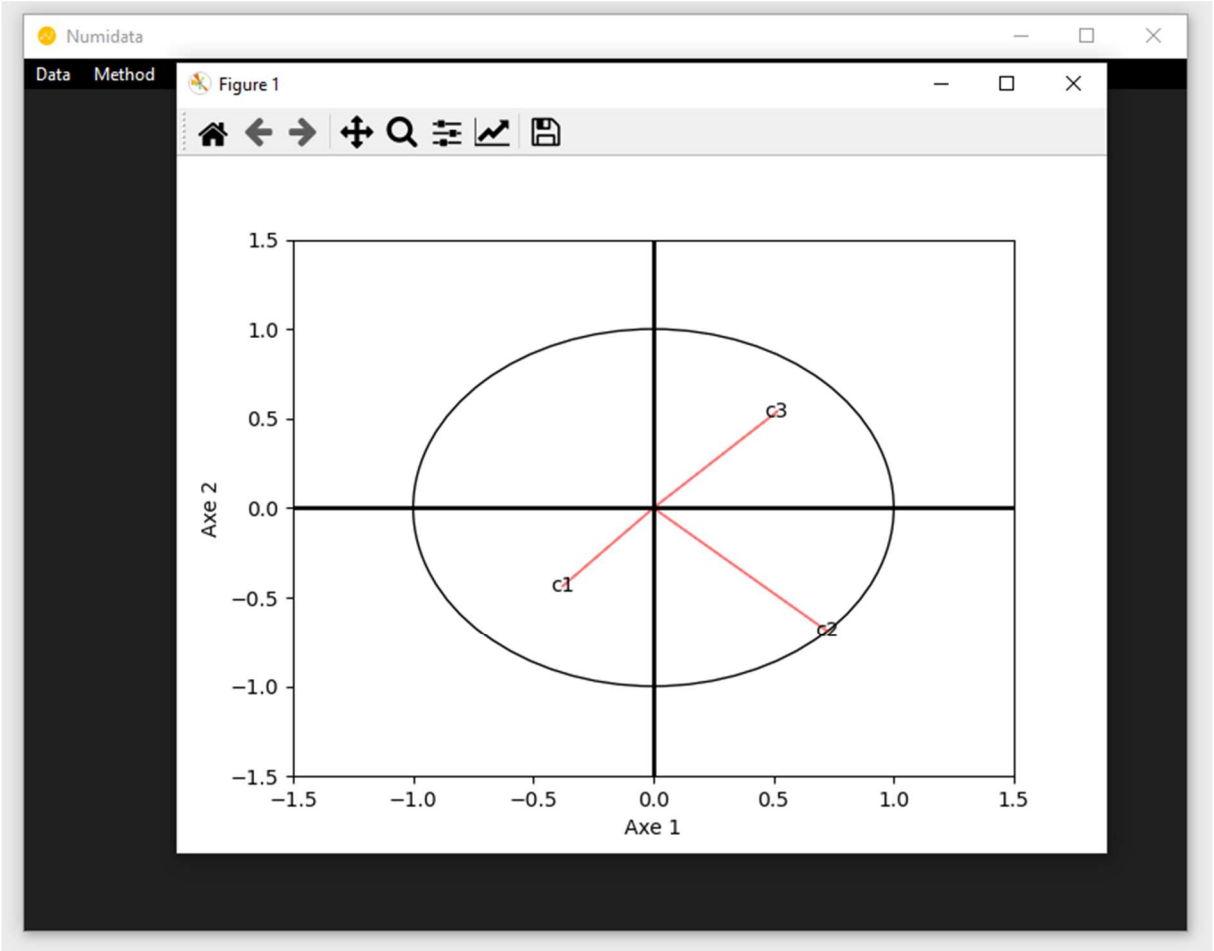

Figure 5.5 : Cercle des corrélations

5.2.2.2. Analyse Factorielle des Correspondances (AFC) :

L'AFC est une méthode d'analyse statistique utilisée pour étudier les relations entre les lignes et les colonnes d'un tableau de contingence. Voici les résultats obtenus lors de l'utilisation de l'AFC :

- Matrice de probabilité : Cette matrice représente la probabilité d'association entre les lignes et les colonnes,
- Matrice DI : Cette matrice représente la diversité des lignes,
- Matrice DJ : Cette matrice représente la diversité des colonnes,
- Matrice LM : Cette matrice représente la liaison mutuelle entre les lignes et les colonnes,
- Matrice CM : Cette matrice représente la concentration des lignes et des colonnes,

- Matrice X_BARRE : Cette matrice représente la matrice des profils lignes moyens centrés,
- Matrice S_BARRE : Cette matrice représente la matrice des profils colonnes moyens centrés,
- Valeurs propres de S_BARRE : Ces valeurs propres indiquent l'importance relative des composantes principales pour les profils colonnes,
- Vecteurs propres de S_BARRE : Ces vecteurs propres indiquent la direction des composantes principales pour les profils colonnes,
- Composantes principales pour les profils lignes : Ces composantes principales représentent la contribution des profils lignes aux valeurs propres,
- Composantes principales pour les profils colonnes : Ces composantes principales représentent la contribution des profils colonnes aux valeurs propres,
- Projection des composantes principales : Ce graphe illustre la relation entre les profils lignes et colonnes dans le plan factoriel.

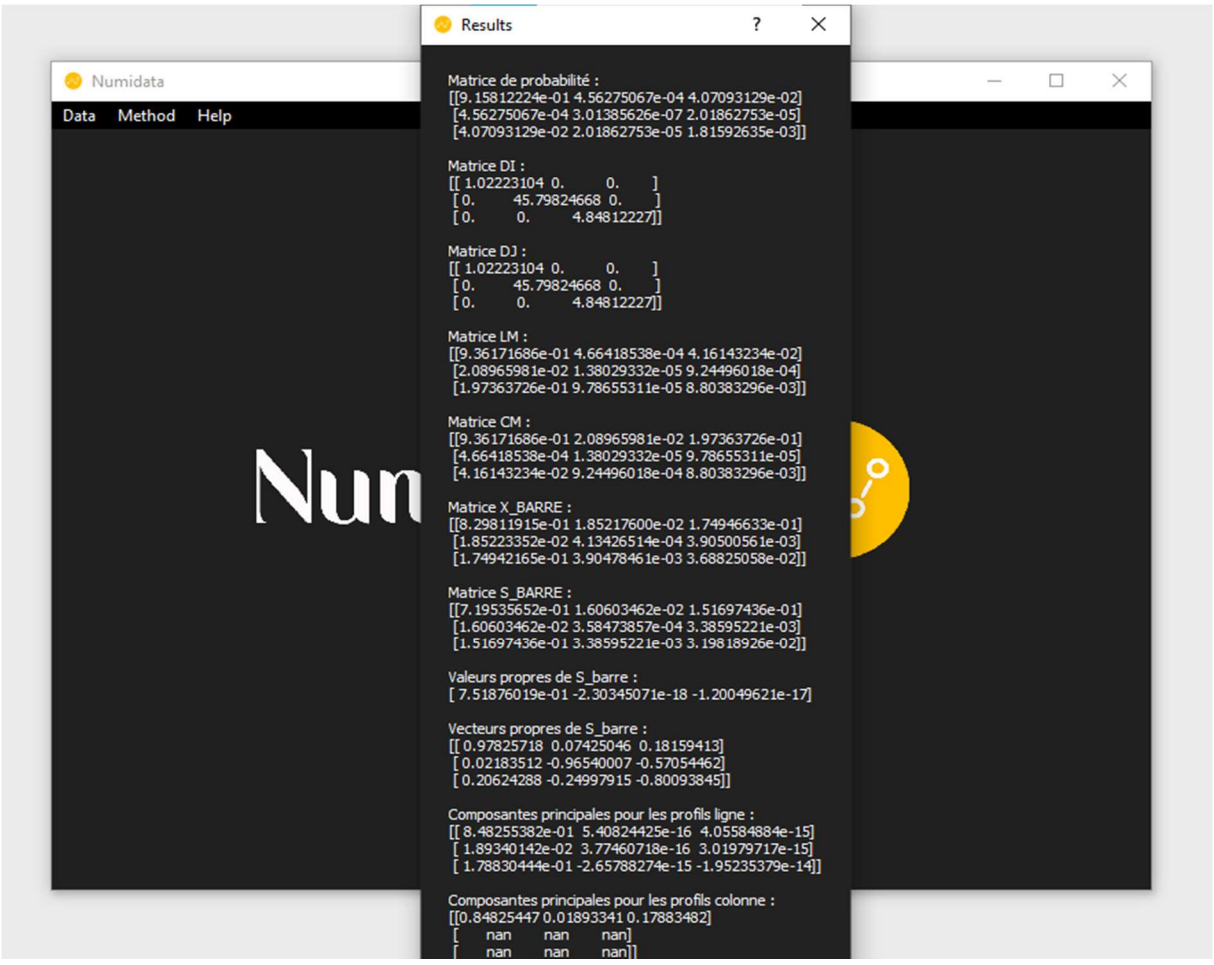

Figure 5.6 : Résultats d'une AFC.

5.2.3. Menu Help

Ce bouton donne accès à des informations supplémentaires sur l'utilisation de l'application. En cliquant sur le bouton "How to", une fenêtre contextuelle s'ouvre, fournissant des instructions détaillées sur l'utilisation de l'application.

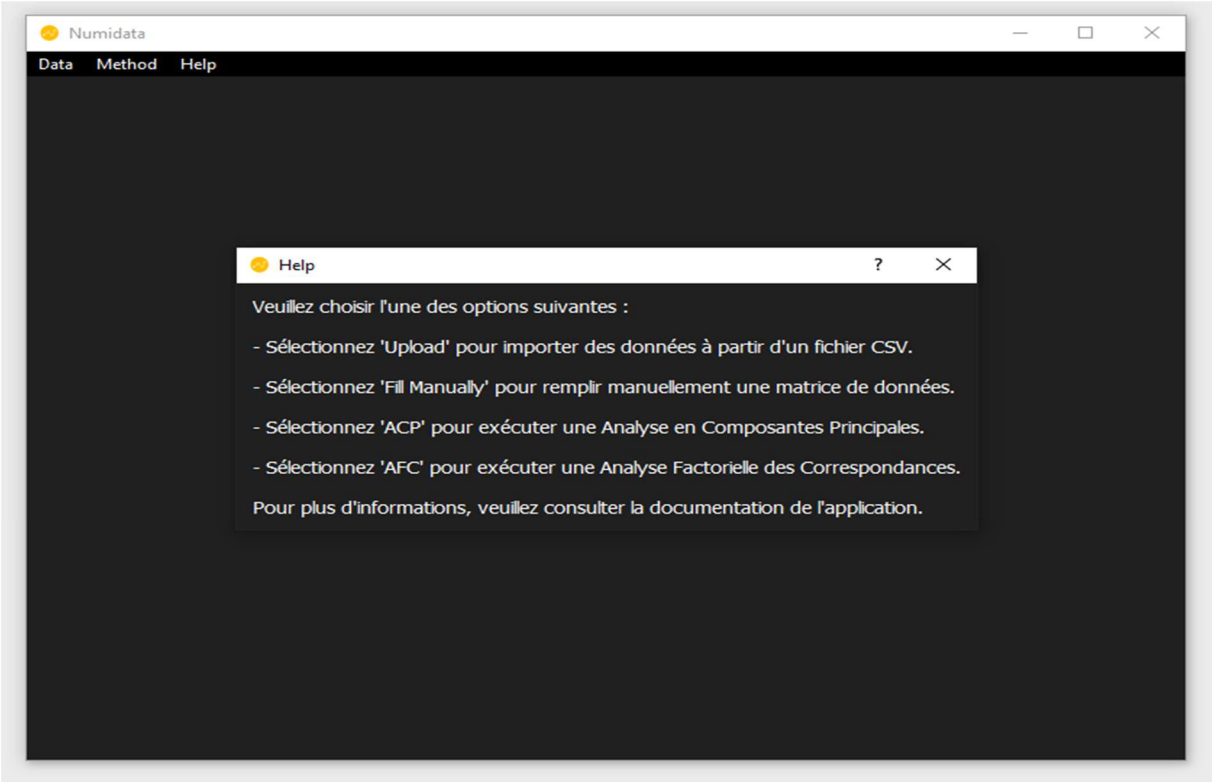

Figure 5.7 : Menu Help.

5.3. Conclusion

En conclusion, ce chapitre a mis en évidence l'importance des logiciels d'analyse des données dans la facilitation et l'interprétation des résultats. L'application NumiData que nous avons développée offre une interface conviviale et des fonctionnalités essentielles pour mener à bien l'analyse en composantes principales (ACP) et l'analyse factorielle des correspondances (AFC). Grâce à cet outil, les utilisateurs ont la possibilité d'effectuer des calculs complexes et de générer des graphiques représentatifs des résultats obtenus.

Conclusion

L'analyse des données, à travers ses différentes techniques, offre des solutions puissantes pour explorer, résumer et interpréter des tableaux de données complexes. Les analyses factorielles, telles que l'analyse en composantes principales (ACP) et l'analyse factorielle des correspondances(AFC), se révèlent être des outils essentiels dans cette démarche.

Au cours de notre travail, nous avons présenté en détail ces deux méthodes, en mettant en évidence leurs spécificités et leurs applications. De plus, nous avons développé une application en langage Python qui facilitera le traitement des données et la visualisation des résultats.

Grâce à cette étude, nous avons pu constater l'importance de l'analyse des données dans la compréhension des relations et des structures sous-jacentes aux données. Les résultats obtenus permettent de prendre des décisions éclairées et d'élaborer des stratégies basées sur une meilleure compréhension des phénomènes étudiés.

Grâce à notre stage pratique au sein de la Division Petroleum Engineering & Development de la SONATRACH, le travail de ce mémoire, nous a offert une précieuse opportunité d'appliquer concrètement la méthode de l'analyse en composantes principales (ACP). Encadrés par Monsieur SOKHAL Abdallah, notre Maître de stage, nous avons pu mettre en pratique nos connaissances théoriques et relever les défis spécifiques au domaine pétrolier. Cette immersion dans l'industrie pétrolière et gazière a renforcé notre conviction quant à l'importance et à la pertinence de l'ACP pour la reconstitution des logs et l'interprétation des données pétro-physiques.

L'ACP s'est révélée être une technique polyvalente et efficace, conférant aux professionnels la capacité d'approfondir leur compréhension des caractéristiques des réservoirs et de prendre des décisions éclairées. Notre expérience pratique a confirmé les avantages substantiels qu'offre cette méthode, en permettant non seulement la reconstitution

des données manquantes ou corrompues, mais également la mise en évidence des relations entre les variables et la détection de tendances significatives.

L'application de l'ACP dans l'interprétation des logs nous a permis de constater ses nombreux avantages. Non seulement elle nous a permis de reconstituer les données manquantes ou corrompues, mais elle nous a également permis d'identifier les relations entre les variables et de détecter des tendances significatives. Cette capacité d'analyse approfondie des données a été précieuse pour comprendre les caractéristiques des réservoirs et optimiser les opérations pétrolières.

L'utilisation de l'application NumiData offre une perspective concrète sur les avantages de l'informatique dans le domaine de l'analyse des données. Les utilisateurs peuvent exploiter les différentes méthodes programmées pour divers domaines d'études et obtenir des informations précieuses à partir de leurs données.

En continuant notre recherche, nous pourrons approfondir nos connaissances et améliorer les fonctionnalités de l'application NumiData. Cela nous permettra d'offrir aux utilisateurs une expérience encore plus enrichissante et de répondre à leurs besoins spécifiques en matière d'analyse des données.

En fin, l'analyse des données joue un rôle essentiel dans de nombreux domaines tels que la recherche scientifique, l'économie, la sociologie, et bien d'autres. Son application, combinée à des outils informatiques modernes, ouvre de nouvelles perspectives pour l'extraction de connaissances à partir de grandes quantités de données.

Annexe A

Les résultats relatifs à l'application de la méthode ACP pour autres puits

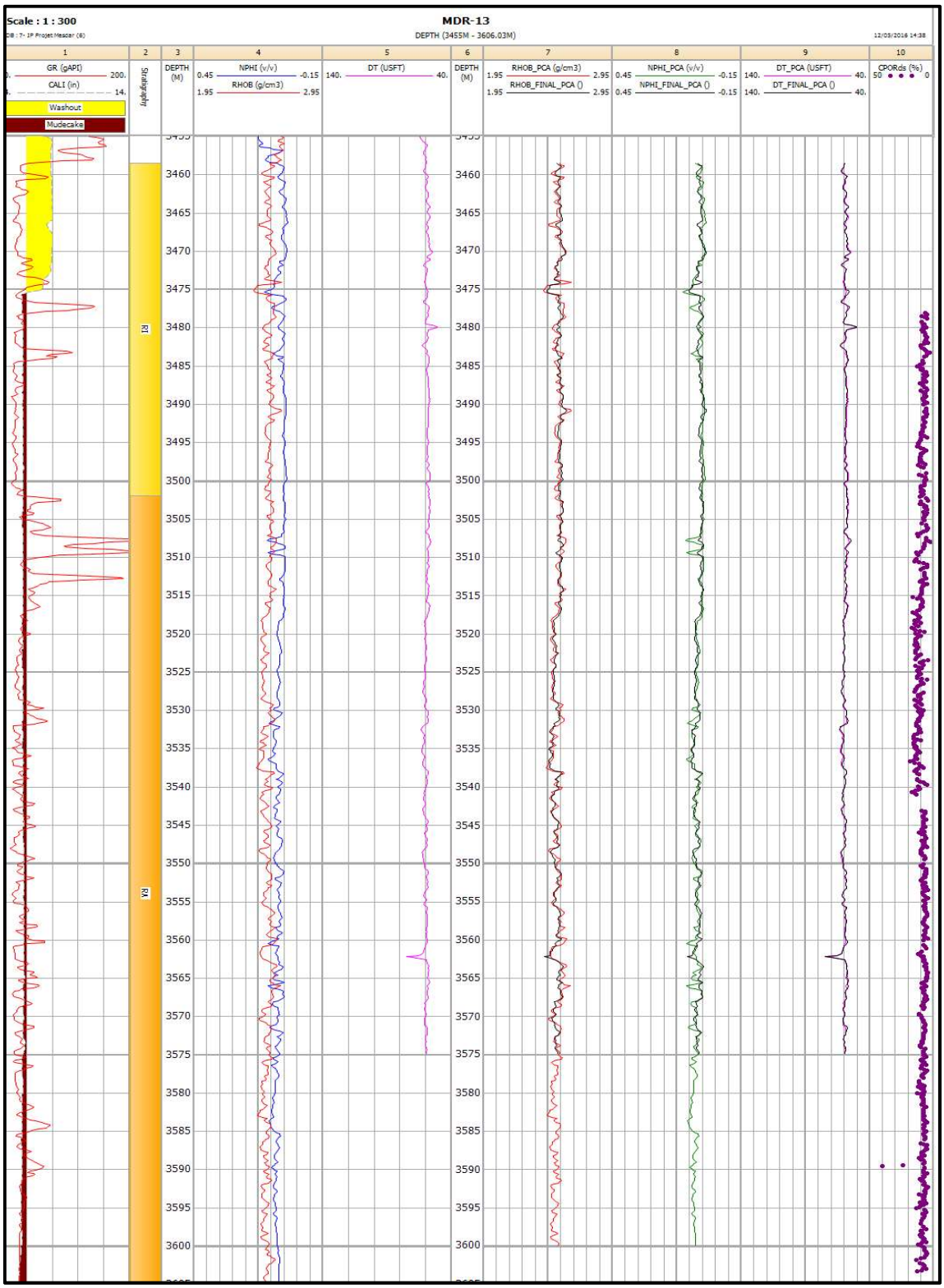

Figure 1 : Well composite du puits MDR-13

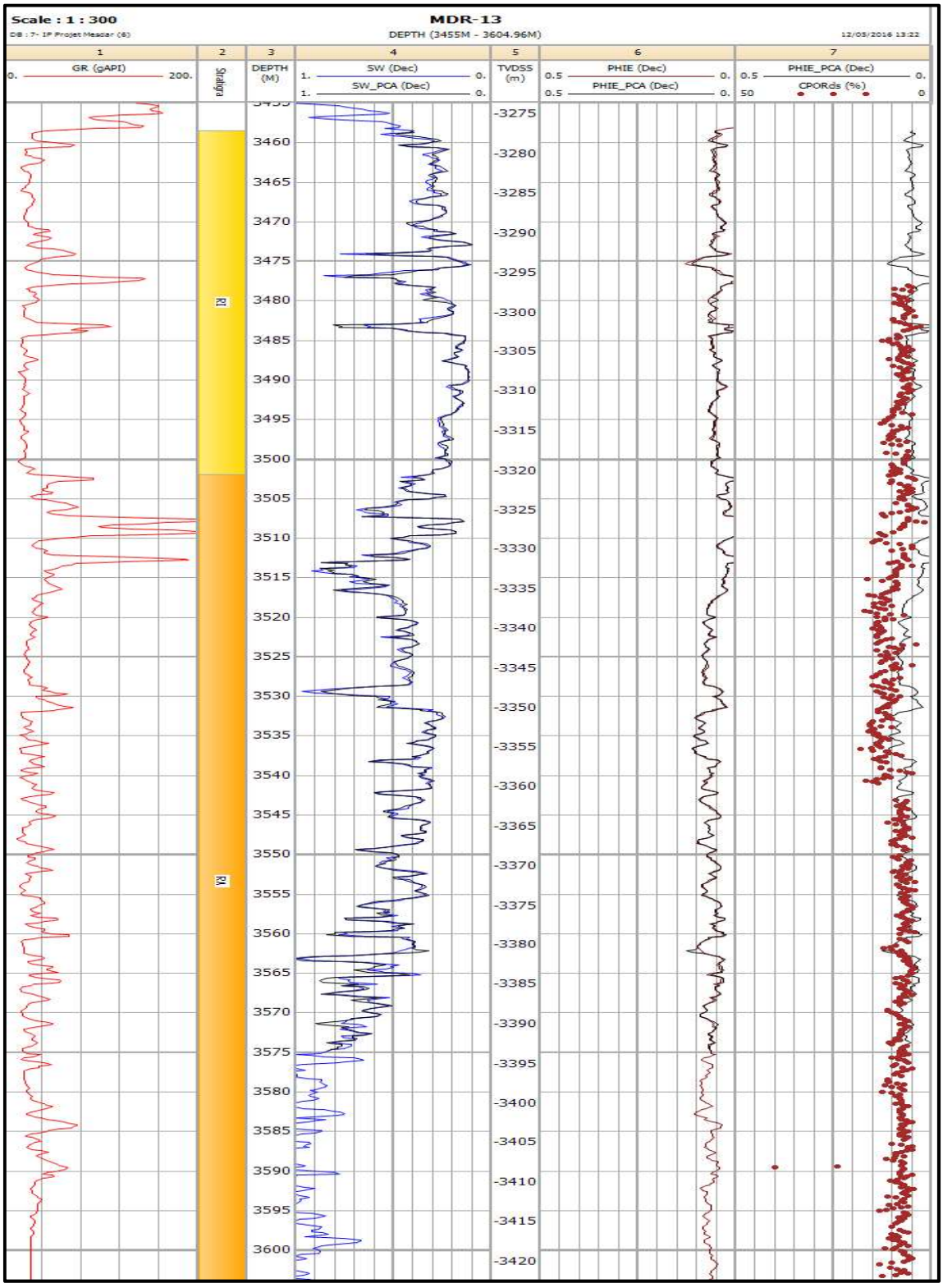

Figure 2 : Réévaluation pétro physique du puits MDR-13 en utilisant les courbes reconstituées par ACP

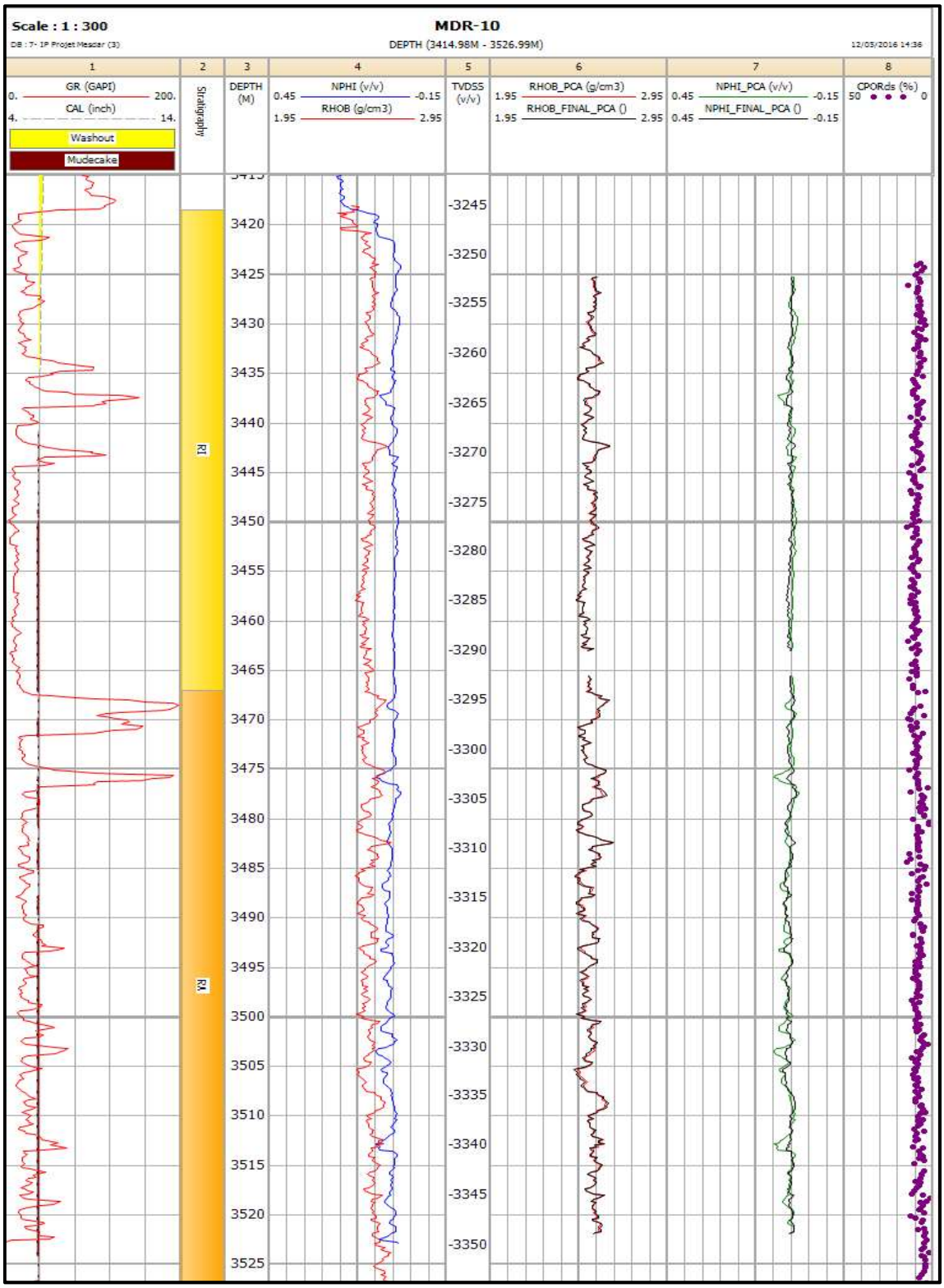

Figure 3 : Well composite du puits MDR-10

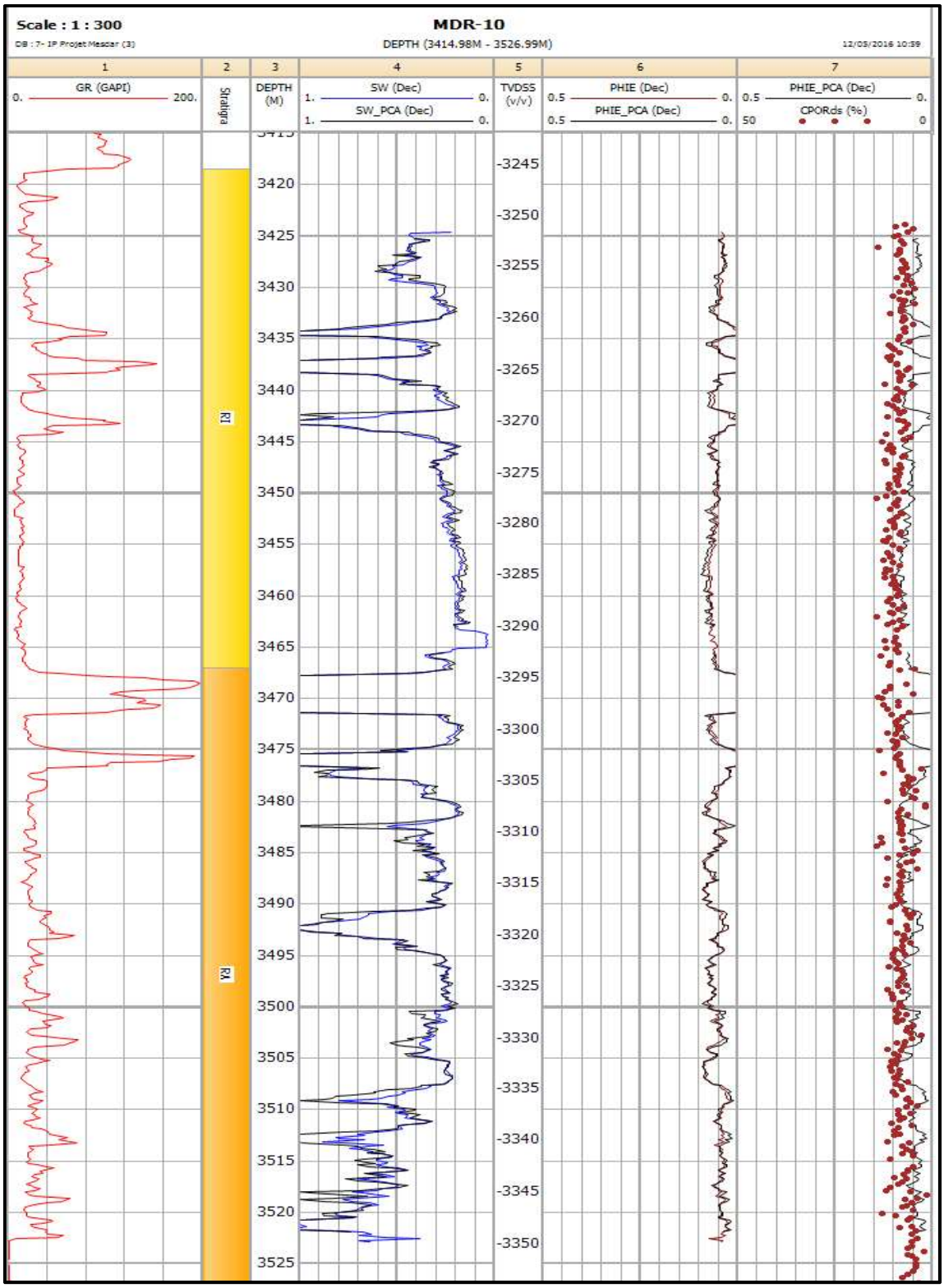

Figure 4 : Réévaluation pétro physique du puits MDR-10 en utilisant les courbes reconstituées par ACP

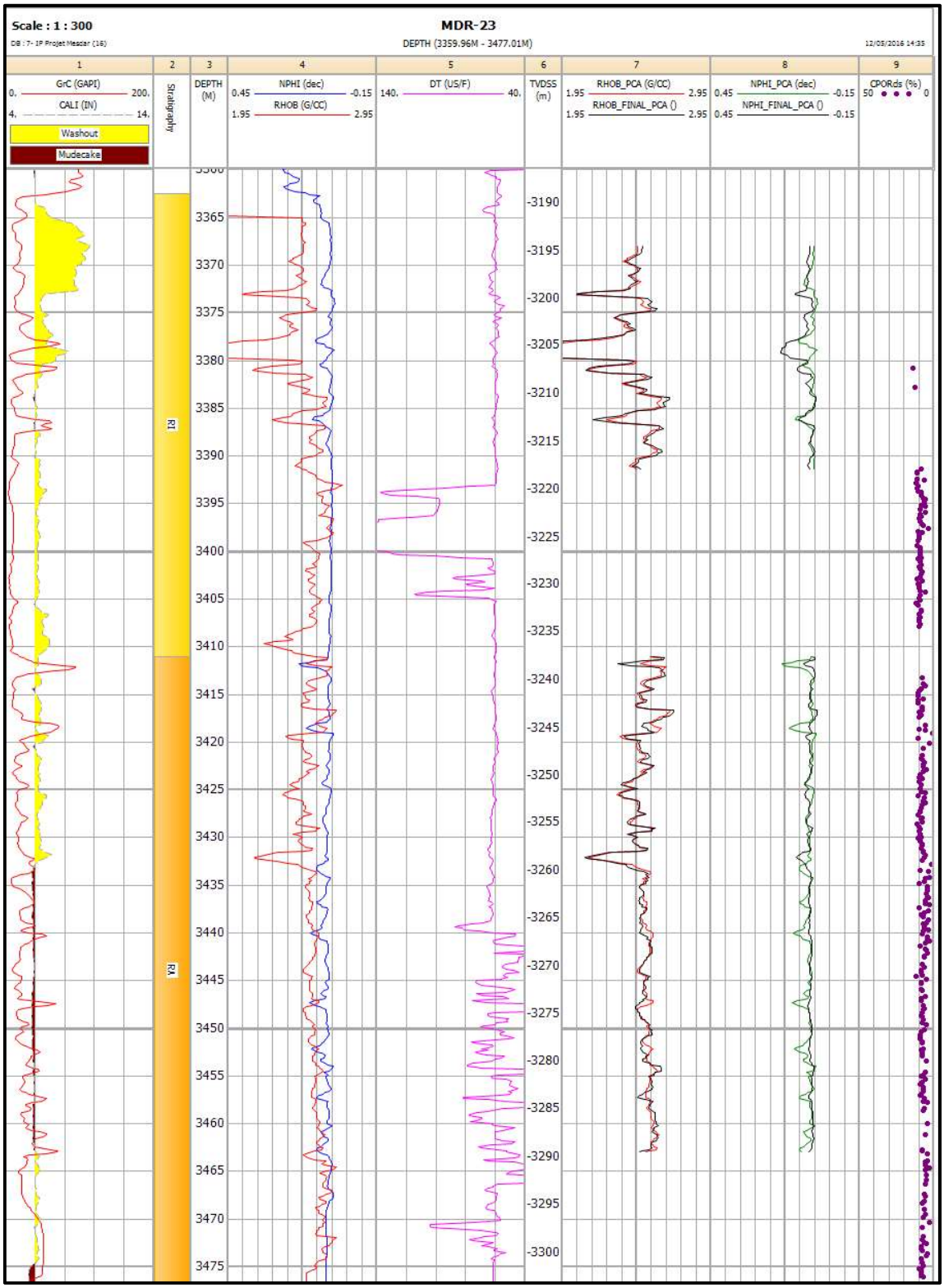

Figure 5 : Well composite du puits MDR-23

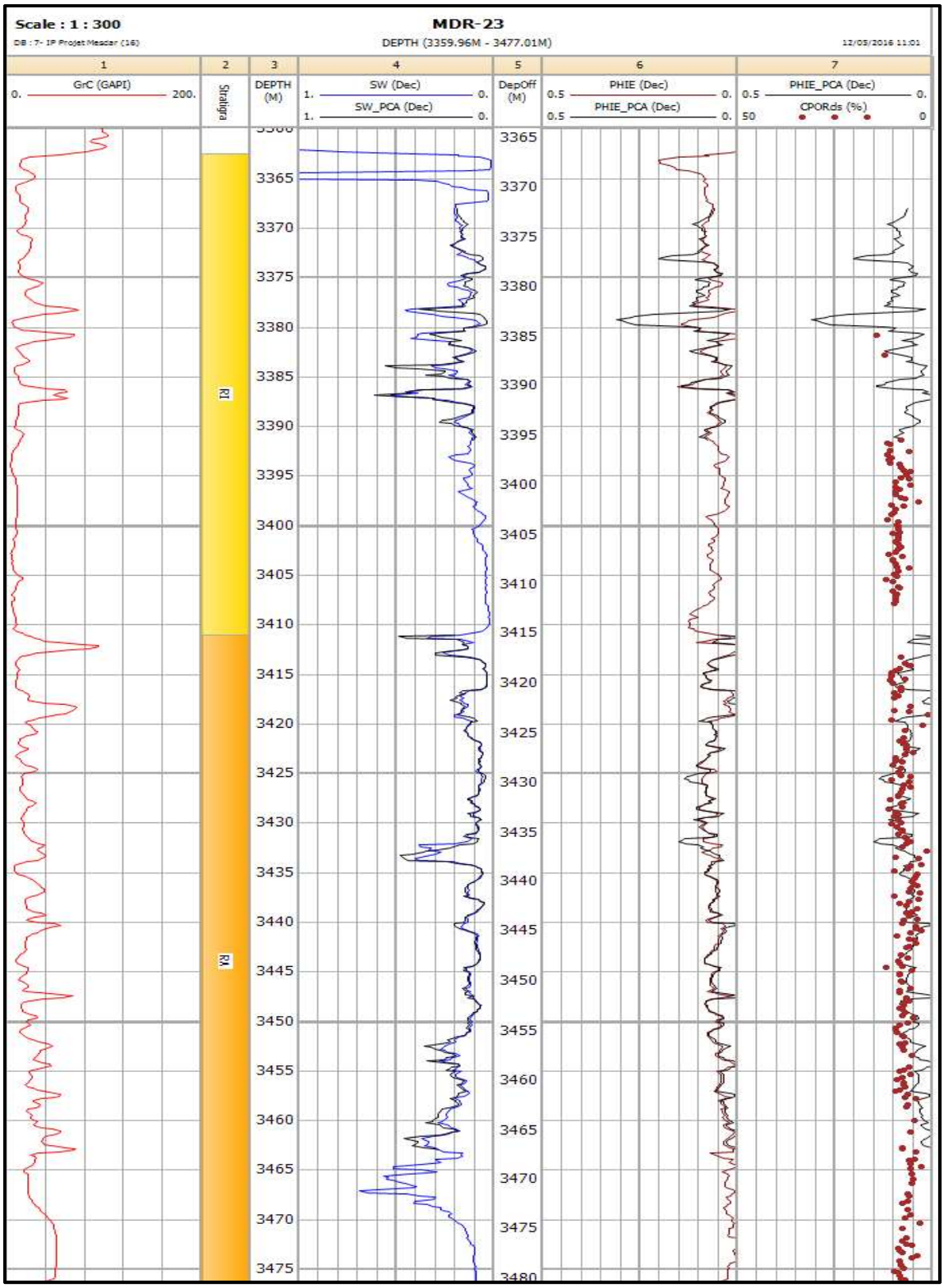

Figure 6 : Réévaluation pétro physique du puits MDR-23 en utilisant les courbes reconstituées par ACP

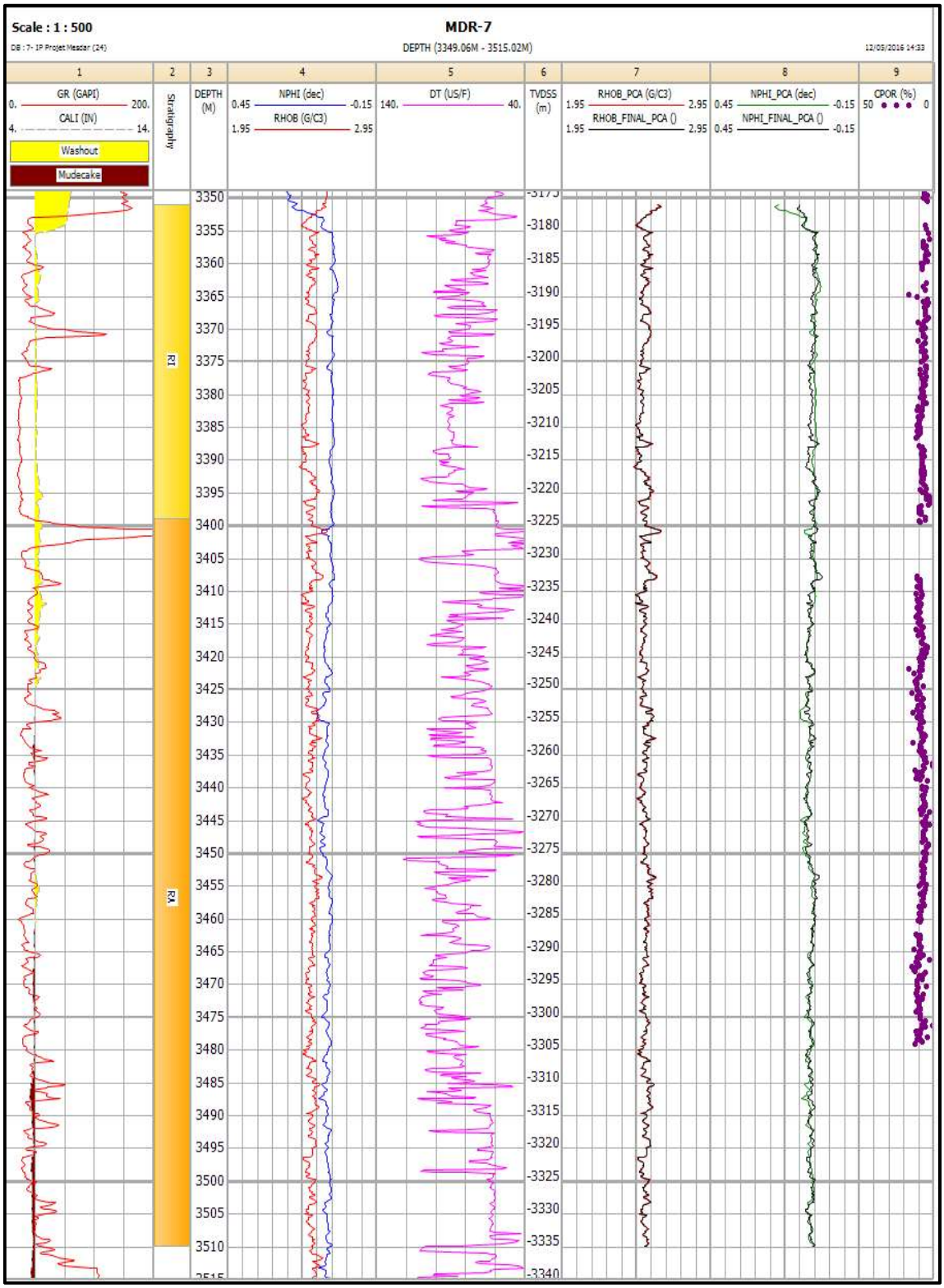

Figure 7 : Well composite du puits MDR-7

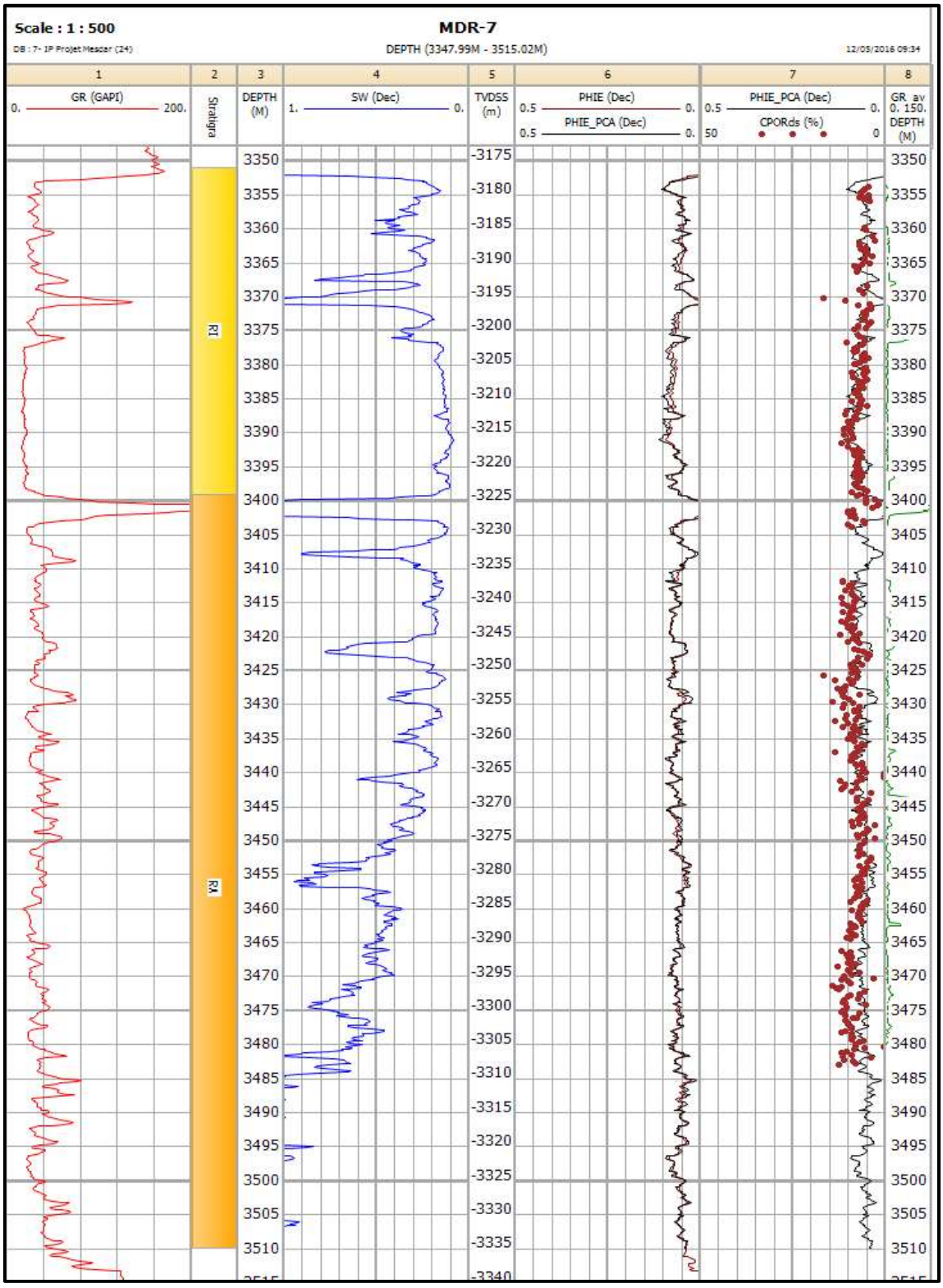

Figure 8 : Réévaluation pétro physique du puits MDR-7 en utilisant les courbes reconstituées par ACP

Références

1. Pages, J.P., Cailliez, F., Escoufier, Y., (1979). Analyse factorielle : un peu d'histoire et de géométrie. Revue de Statistique Appliquée, Vol XXVII, n°1 pp. 5-28.
2. Lebart, L., Morineau, A., Tabard, N., (1977). Techniques de la description statistique. Méthodes et logiciels pour l'analyse de grands tableaux, Dunod.
3. Benzécri, J.-P., (1982). Histoire et préhistoire de l'analyse des données, Dunod.
4. Bouroche, J.M., (1977). Analyse des données en marketing, Masson.
5. Fenelon, J.P., (1981). Qu'est-ce que l'analyse des données ?, Lefonen.
6. Escofier, B., Pages, J., (1988). Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation, Dunod.
7. Gibrat, R., (1978). L'analyse des données : Première partie : Journal de la Société de Statistique de Paris n°3, pp.201-228. Deuxième partie : les sciences humaines : impasse, échecs et succès. Journal de la Société de statistique de Paris n°4 pp.312-331.
8. Lebart, L., Morineau, A., Tabard, N., (1977). Techniques de la description statistique. Méthodes et logiciels pour l'analyse de grands tableaux, Dunod.
9. Saporta, G. (2010). Probabilités, analyse des données et statistiques, 2ème Èdition. Technip, Dunod.
10. Cailliez, F. Pages J.P., (1976). Introduction à l'analyse des données, Smash.
11. Nakache, J.P., Chevalier, A., Morice, V., (1981). Exercices commentés de mathématiques pour l'analyse statistique des données, Dunod.
12. Escofier-Cordier, B., (1965). L'analyse factorielle des correspondances. Cahiers du Bureau universitaire de recherche opérationnelle, n°13.

13. Escofier, B., Pages, J., (1988). Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation, Dunod.

14. Dervin, C., (1990). Comment interpréter les résultats d'une analyse factorielle des correspondances. ITCF.