

UNIVERSITE KASDI MERBAH – OUARGLA

FACULTÉ DES HYDROCARBURES, DES ÉNERGIES RENOUVELABLES ET DES
SCIENCES DE LA TERRE ET DE L'UNIVERS

DEPARTEMENT DES SCIENCES DE LA TERRE ET DE L'UNIVERS



Mémoire de Master Académique

Domaine : Sciences de la Terre et de l'Univers

Filière : Géologie

Spécialité : Géologie Pétrolière

THEME

**Contribution de (Feature Engineering) à l'amélioration des Modèles de
la Régression Linéaire Multiple pour la Prédiction du Carbone
Organique Totale, Application aux Schistes de champs de Barnett
(Etats Unis)**

Présenté par

HABIRECHE NADIR

Soutenu publiquement le 26-06-2019

Devant le jury :

Président :	MELOUAH Oualide	MCB	Univ. Ouargla
Promoteur :	AMEUR ZAIMECHE Ouafi	MAA	Univ. Ouargla
Examineur :	ZATOUT Merzouk	MCB	Univ. Ouargla

Année Universitaire : 2018/2019

Remerciements

Je voudrais remercier en premier **ALLAH** le tout puissant de m'avoir donné la puissance et la force nécessaire pour achever ce travail.

Ce travail a été dirigé par **Mr. AMEUR ZAIMECHE Ouafi**, mon Promoteur Qu'il trouve ici le témoignage de mes profondes gratitude pour ses conseils, ses remarques et sa disponibilité tout au long de cette étude. Nous tenons à lui exprimer notre profonde reconnaissance pour la patience et la bienveillance dont elle a fait preuve.

Un grand merci aux membres de jury qui ont accepté d'examiner ce Travail **Dr. MELOUAH Oualide** Qui m'a fait l'honneur de présider le jury de soutenance

Dr. ZATOUT Merzouk d'avoir accepté d'examiner ce travail

Je remercie mes très **chers parents** qui ont toujours été là pour moi, Vous avez tout sacrifié pour vos enfants n'épargnant ni santé ni efforts. Vous m'avez donné un magnifique modèle de labeur et de persévérance. Je suis redevable d'une éducation dont je suis fier .

Je voudrais remercier, à travers ce mémoire, tous les enseignants du département Sciences de la Terre et de l'Univers 2018/2019 qui nous ont constamment été d'un immense soutien tout au long de ce travail et j'adresse une pensée particulière, à l'enseignant de spécialité Géologie Pétrolier

Je voudrais présenter mes sincères remerciements les plus distingués à toute personne qui a contribué à cette formation.

Nadir

Dédicaces

Je dédie ce travail qui n'aura jamais pu voir le jour sans les soutiens indéfectibles et sans limite de mes chers parents qui ne cessent de me donner avec amour le nécessaire pour que je puisse arriver à ce que je suis aujourd'hui. Que dieux vous protège et que la réussite soit toujours à ma portée pour que je puisse vous combler de bonheur.

Je dédie aussi ce travail à :

Ma femme et sa famille. Mes frères, mes sœurs et leur famille. Mes oncles, mes tantes et leur famille. Tous mes cousins et cousines. Tous mes amis, mes collègues et tous ceux qui m'estiment.

NADIR

SOMMAIRE

Remercîment	I
Dédicace	II
Résumé	III
Liste des figures	IV
Liste des tableaux	V
Abréviations	VI
Introduction générale et La problématique.....	1

Chapitre I : Généralité sur les gisements non conventionnels

1-Les gisements conventionnels et non conventionnels	3
1.1-Classification des hydrocarbures non conventionnels liquides	5
1.2-Classification des hydrocarbures non conventionnels gazeux	6
1.3-Les hydrocarbures de roche-mère	7
2-Etude d'un cas de schistes de Barnett (ou Barnett shale)	9
2.1-Contexte géologique et stratigraphique du bassin de Fort Worth	9
3-Généralités sur le carbone organique totale COT	11
3.1-Origines de la matière organique	12
3.2-Préservation de la matière organique	13
3.3-Maturation de la matière organique	13

Chapitre II : Matériel et Méthodes

1-Les diagraphies utilisées (well logging).....	15
1.1-But des diagraphies.....	15
1.2-Classification des diagraphies.....	15
1.2.1. Diagraphies nucléaires.....	16
1.2.2. Diagraphies électriques.....	17
1.2.3. Diagraphie sonique (ondes acoustiques).....	18

2-Préparation de la base de données	19
3-Méthodologie	21
3.1-Mesure directe	21
3.2-Méthodes indirectes	22
3.2.1-La Régression Multiple	22
3.2.2. Feature Engineering (Transformation des données).....	23
3.2.3. D'Élimination des Outliers.....	25
3.2.4. La Validation externe.....	25
4- Les indice de performance.....	26
5-LOGICIELS UTILISES.....	27
5.1-Logiciel "GetData Graph Digitizer".....	27
5.2-Logiciel "EXCEL".....	27
5.3-Logiciel "STATISTICA".....	27
Conclusion.....	28

Chapitre III : Résultats et discussion

1- Analyse des données	29
2- Prédiction de COT par la méthode de Régression Multiple en différents modèles.....	30
2.1 Prédiction de COT avec les data brut (sans transformation).....	30
2.2 Prédiction de COT avec transformation de variables.....	32
2.3 Prédiction de COT transformé avec transformation des DATA.....	38
2.4 D'Élimination des OUTLIERS.....	45
2.4 La validation Externe.....	50
Conclusion	52

Conclusion générale

PERSPECTIVES

Références bibliographiques

Annexs

Liste des figures

Figure .1 : Un gisement des hydrocarbures conventionnel.....	4
Figure .2 : Les hydrocarbures non conventionnels.....	4
Figure .3 : Le différent type d'hydrocarbures liquides non conventionnels.....	5
Figure .4 : Les différents types d'hydrocarbures gazeux non conventionnels.....	6
Figure .5 : Les hydrocarbures de roche-mère.....	8
Figure .6 : Log lithologique avec le COT% en poids correspondant, pour le schiste De Barnett.....	10
Figure .7 : Carte de localisation du Barnett shale-bassin de Fort Worth.....	11
Figure .8 : Roche mère de type shale (argileuse feuilletée), contenant de la matière organique.....	12
Figure .9 : Schéma des différentes diagraphies.....	16
Figure .10 : Une illustration montrant les différentes variables liées à la résistivité.....	18
Figure .11 : principe de mesure du Sonique.....	19
Figure .12 : Conversion de données graphiques en données numériques.....	20
Figure .13 : photo de carotte du puits.....	21
Figure .14 : Dispositif de mesure du carbone organique total.....	22
Fig. 15 : Organigramme explicatif de la prédiction du carbone organique total (COT)....	28
Fig. 16 : Comparaison graphique entre le COT mesuré et estimé par la méthode de RLM utilisant les data brut (sans transformation de variables).....	30
Fig. 17 : Comparaison graphique entre le 1/COT mesuré et estimé par la méthode de RLM utilisant les data brut.....	32
Fig.18 : Comparaison graphique entre le LOG ₁₀ (COT) mesuré et estimé par la méthode de RLM utilisant les data brut.....	33
Fig. 19 : Comparaison graphique entre le EXP(COT) mesuré et estimé par la méthode de RLM utilisant les data brut.....	35
Fig. 20 : Comparaison graphique entre la Racine(COT) mesuré et estimé par la méthode de RLM utilisant les data brut (sans transformation de variables).....	36
Fig. 21 : Comparaison graphique entre le 1/TOC ² mesuré et estimé par la méthode de RLM utilisant les data brut.....	37

Fig. 22 : Comparaison graphique entre le $1/COT$ mesuré et estimé par la méthode de RLM utilisant l'inverse des DATA.....	39
Fig. 23 : Comparaison graphique entre le $1/COT^2$ mesuré et estimé par la méthode de RLM utilisant l'inverse ² des data.....	40
Fig. 24 : Comparaison graphique entre le $LOG_{10}(COT)$ mesuré et estimé par la méthode de RLM utilisant $LOG_{10}(data)$	41
Fig. 25 : Comparaison graphique entre la Racine(COT) mesuré et estimé par la méthode de RLM utilisant la Racine de data.....	42
Fig. 26 : Comparaison graphique entre le $1/COT$ mesuré et estimé par la méthode de RLM utilisant les data brut après la d'élimination des outliers.....	45
Fig. 27 : Comparaison graphique entre le $LOG_{10}(COT)$ mesuré et estimé par la méthode de RLM utilisant les data brut après la d'élimination des outliers.....	46
Fig. 28 : Comparaison graphique entre le $LOG_{10}(COT)$ mesuré et estimé par la méthode de RLM utilisant le $LOG_{10}(data)$ après la d'élimination des outliers.....	47
Fig. 29 : Comparaison graphique entre la Racine(COT) mesuré et estimé par la méthode de RLM utilisant la Racine de data après la d'élimination des outliers.....	48

Liste des tableaux

Tab .1 : Teneur en COT définissant le potentiel pétrolier d'une roche mère.....	12
Tab.02: Tableau statistiques descriptive de COT mesuré et les enregistrements diagraphiques.....	29
Tab. 03 : Matrice de corrélation de COT mesuré et les enregistrements Diagraphiques.....	30
Tab. 04. Illustration de comparaison statistique entre le COT mesuré et estimé par la méthode de la RLM utilisant les data brut (sans transformation de variables).....	31
Tab. 05. Illustration de comparaison statistique entre le 1/COT mesuré et estimé par la méthode de la RLM utilisant les data brut.....	32
Tab. 06. Illustration de comparaison statistique entre $\text{LOG}_{10}(\text{COT})$ mesuré et estimé par la méthode de la RLM utilisant les data brut.....	34
Tab. 07. Illustration de comparaison statistique entre $\text{EXP}(\text{COT})$ mesuré et estimé par la méthode de la RLM utilisant les data brut.....	35
Tab. 08. Illustration de comparaison statistique entre la Racine(COT) mesuré et estimé par la méthode de la RLM utilisant les data brut.....	36
Tab. 09. Illustration de comparaison statistique entre $1/\text{TOC}^2$ mesuré et estimé par la méthode de la RLM utilisant les data brut.....	37
Tab. 10. Illustration de comparaison statistique entre 1/COT mesuré et estimé par la méthode de la RLM utilisant l'inverse des DATA.....	39
Tab. 11. Illustration de comparaison statistique entre $1/\text{COT}^2$ mesuré et estimé par la méthode de la RLM utilisant l'inverse ² des data.....	40
Tab. 12. Illustration de comparaison statistique entre $\text{LOG}_{10}(\text{COT})$ mesuré et estimé par la méthode de la RLM utilisant $\text{LOG}_{10}(\text{data})$	41

Tab. 13. Illustration de comparaison statistique entre la Racine(COT) mesuré et estimé par la méthode de la RLM utilisant la Racine de data.....	43
Tab. 14. comparaison statistique entre les différents modèles de la RLM.....	44
Tab. 15. Illustration de comparaison statistique entre 1/COT mesuré et estimé par la méthode de la RLM utilisant les data brut après la d'élimination des outliers.....	45
Tab. 16. Illustration de comparaison statistique entre LOG ₁₀ (COT) mesuré et estimé par la méthode de la RLM utilisant les data brut après la d'élimination des outliers.....	46
Tab. 17. Illustration de comparaison statistique entre LOG ₁₀ (COT) mesuré et estimé par la méthode de la RLM utilisant le LOG ₁₀ (data) après la d'élimination des outliers.....	48
Tab. 18. Illustration de comparaison statistique entre la Racine(COT) mesuré et estimé par la méthode de la RLM utilisant la Racine de data après la d'élimination des outliers.....	49
Tab. 19. comparaison statistique entre les différents modèles de la RLM.....	49
Tab. 20. comparaison statistique entre les coefficients de déterminations de 11 validations externes.....	51

ABREVIATIONS

COT	CARBONE ORGANIQUE TOTAL
LMR	régression linière multiple
CC	COEFFICIENT DE CORRELATION
REMA	EREUR QUADRATIQUE MOYENNE
MAE	L'EREUR ABSOLUE MOYEN
V D	Variable dépendant
V Id	Variable indépendant

INTRODUCTION GENERALE

Introduction Générale et problématique

L'évaluation précise des propriétés de la roche mère est primordiale pour l'exploration de pétrole et de gaz. En tant qu'indice fondamental et important, la teneur en carbone organique total (COT) peut représenter l'abondance de la matière organique et la fertilité de la roche. Ce paramètre est généralement obtenu à partir des analyses de laboratoire qui sont coûteuses et prennent beaucoup de temps. Elles se limitent le développement rapide de l'exploration pétrolière et gazière non conventionnelle. Étant donné que de nombreuses méthodes géophysiques ont été utilisées pour déterminer les paramètres de formation (*Li et Wang, 2016; Wang et al. 2016*), il est possible d'estimer le contenu en COT à partir de données de diagraphies.

Les chercheurs ont réalisé de nombreuses équations pour améliorer continue les prévisions de COT, mais dépend de la disponibilité des carottes, la récupération n'est pas toujours intégrale à cause des problèmes techniques et géologiques rencontrés durant l'opération du carottage (*Ameur Zaimeche et al 2014*).

Elle existe une relation entre la teneur en COT et certaines types de diagraphies tels que ; le neutron, le gamma naturel, la densité, la résistivité et le sonic. *Beers et al. (1945) et Schmoker J (1979)* ont calculé le contenu en COT utilisant le gamma ray, qui s'est avérée appropriée pour la roche mère riche en éléments radioactifs. *Schmoker et Hester (1983), Meyer and Nederlof (1984), and Decker et al. (1993)* ont calculé le contenu de COT en utilisant la densité. En conséquence, ils n'ont pu obtenir des meilleures prévisions que lorsqu'il existe une forte corrélation entre le log de densité et le contenu du COT. *Autric and Dumesnil (1985)* ont calculé le contenu du COT en utilisant le log sonique ce qui permis une bonne prédiction lorsqu'il existe une forte corrélation entre log sonique et le contenu du COT.

Les équations empiriques ne répondent pas à la prédiction du contenu du COT à l'aide des diagraphies, à cause de l'influence des différences paramètres physiques dans la zone d'étude, et il a été constaté que l'utilisation d'une formule empirique était défavorable à la bonne prédiction du contenu du COT *Guo et al., (2009), Hu et al. (2011), Huo et al., (2011)* ont calculé la teneur en COT en utilisant la corrélation entre les diagraphies des puits de résistivité et de porosité neutron. Les résultats ont prouvé sa simplicité et restent très limités.

Afin d'améliorer la prédiction du contenu du COT à l'aide des diagraphies de puits, *Heidari et al. (2011)* ont sélectionné plusieurs logs afin d'établir une équation de régression linéaire multiple pour la prédiction du contenu de la COT. Cependant, il s'est avéré difficile de se renseigner sur les paramètres complexes en raison de la relation non linéaire entre les différents enregistrements de diagraphies.

Alors dans notre présent mémoire on examine différents modèles de la régression multiple, notamment le Feature Engineering (transformations des données), la réduction des outliers et la validation externe pour la prévision du COT dans le champ de Barnett. Et pour évaluer l'efficacité de ces modèles une comparaison qualitative (graphique) et quantitative les indices de performances numériques) entre les valeurs observées et les valeurs simulées du TOC.

La problématique

En réalité la meilleur méthode d'estimation du COT est la mesure direct au l'laboratoire, mais elle est couteuse et prend du temps. Pour cela les chercheurs ont développé des équitations pour l'estimé, dans notre travail on essaye d'améliorer les performances de la régression linière multiple pour estimer le carbone organique total dans les intervalles non carottées sur le Barnett.

Alors quel sont les modèles qui perdirent les lacunes du COT, minimisent le cout et gain du temps, et donnent la moindre d'erreur ??

Chapitre I

Généralité sur les gisements non conventionnels

Introduction

Avant de parler de gisements d'hydrocarbures non conventionnels, nous allons donner quelques informations sur les gisements dits conventionnels en faisant ressortir leurs caractéristiques.

Il est très important de comprendre les caractéristiques qui nous permettent de distinguer un gisement conventionnel d'un autre non conventionnel. Physiquement et chimiquement, rien ne différencie un hydrocarbure non conventionnel d'un hydrocarbure conventionnel, il s'agit toujours de pétrole (hydrocarbure liquide) ou de gaz (gaz naturel, principalement composé de méthane), le classement d'un hydrocarbure dans l'une ou l'autre catégorie tient aux conditions d'accumulation de cet hydrocarbure dans le sous-sol et aux types de technologies à mettre en œuvre pour l'en extraire (Rouaud, 2013).

1. Les gisements conventionnels et non conventionnels

L'enfouissement et la compaction de multiples couches de matières organiques à des températures et des pressions de plus en plus importantes, il y a plusieurs centaines de millions d'années, a permis la fabrication de pétrole et de gaz au sein de la roche (Théorêt, 2015).

La génération des hydrocarbures (conventionnel et non conventionnel) qui écrit ci-dessus à été faite dans la roche mère:

La première, soit un gisement conventionnel. Il faut que quatre conditions soient réunies dans une même région :

une couche riche en matière organique (la roche-mère) qui va, par augmentation de pression et de température, se transformer en hydrocarbures, lorsque la température et la pression augmentent, la matière organique se transforme d'abord en pétrole puis en gaz (principalement du méthane) ;

une couche poreuse et perméable (le réservoir) dans laquelle les hydrocarbures vont se concentrer et pouvoir être produits ;

une couche imperméable (la couverture) qui empêche la migration des hydrocarbures vers la surface ;

un piège dans lequel les hydrocarbures vont se concentrer ;

Les hydrocarbures se trouvant dans une roche poreuse et perméable et concentrés dans un piège (gisement), ils sont relativement faciles à produire (Ifpen, 2012).

Un forage vertical jusqu'à la roche-réservoir permet aujourd'hui de faire jaillir le pétrole ou le gaz sous pression (Théorêt, 2015).

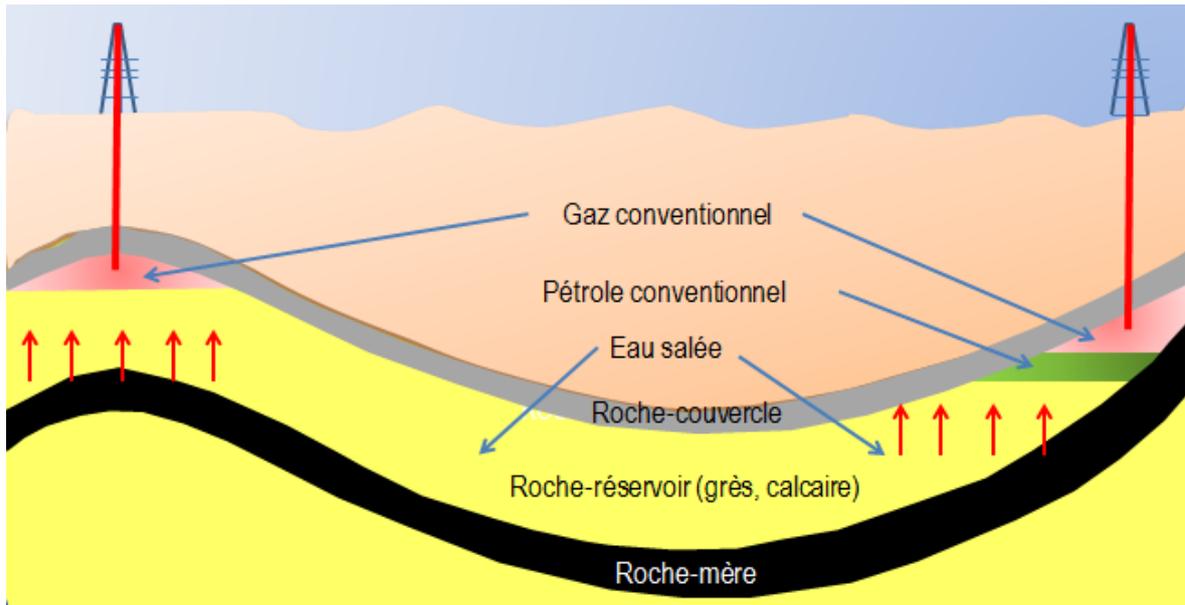


Fig. 01 : Un gisement des hydrocarbures conventionnel
(US Geological Survey, no date)

- La seconde, soit un **gisement non conventionnel**, requérait également une roche mère, mais pas de roche-réservoir ni de roche couverture, En effet, la roche mère maintient les hydrocarbures emprisonnés, étant donné son faible degré de porosité et de perméabilité (Théorêt, 2015).

Afin d'accéder aux hydrocarbures maintenus dans la roche, et considérant les connaissances scientifiques actuelles, il sera presque toujours nécessaire de procéder à la fracturation de la roche (Théorêt, 2015).

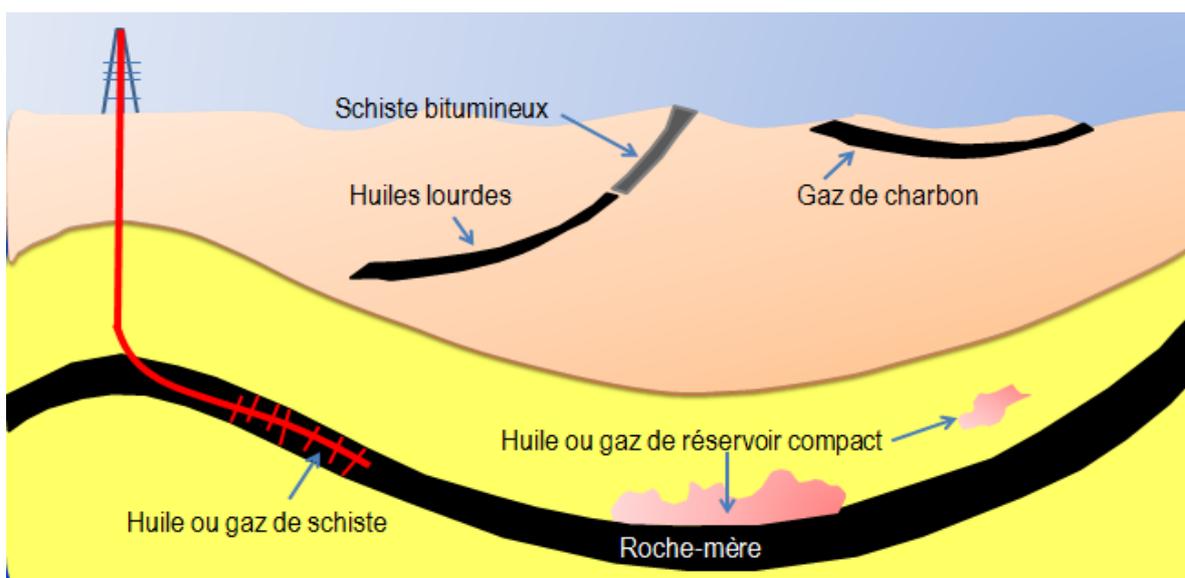


Fig. 02 : Les hydrocarbures non conventionnels(US Geological Survey, no date)

Sous le nom générique d'hydrocarbures non conventionnels, on distingue de nombreux types des hydrocarbures, tant liquides que gazeux, Ils sont présentés ci-après, selon une représentation sous forme de triangle classiquement utilisée pour la classification des hydrocarbures en fonction de la nature de la roche dans laquelle ils sont piégés (Rouaud, 2013).

Classification des hydrocarbures non conventionnels liquides

La pointe supérieure du triangle (en jaune) représente les hydrocarbures liquides conventionnels piégés dans des réservoirs de bonne qualité, faciles à développer, Au-dessous (en orange), on trouve les hydrocarbures non conventionnels piégés dans de très mauvais réservoirs (pétroles de réservoirs compacts, "tight oil") ou des pétroles visqueux (pétroles lourds ou extra-lourds "heavy oils"), voire des bitumes comme les sables bitumineux ("oil sands"),

La partie basse (en violet) regroupe les hydrocarbures liquides non conventionnels piégés dans la roche mère: les schistes bitumineux, dans lesquels la matière organique n'a pas été suffisamment transformée pour générer des hydrocarbures, et les pétroles de schistes, non expulsés de la roche-mère (Rouaud, 2013).

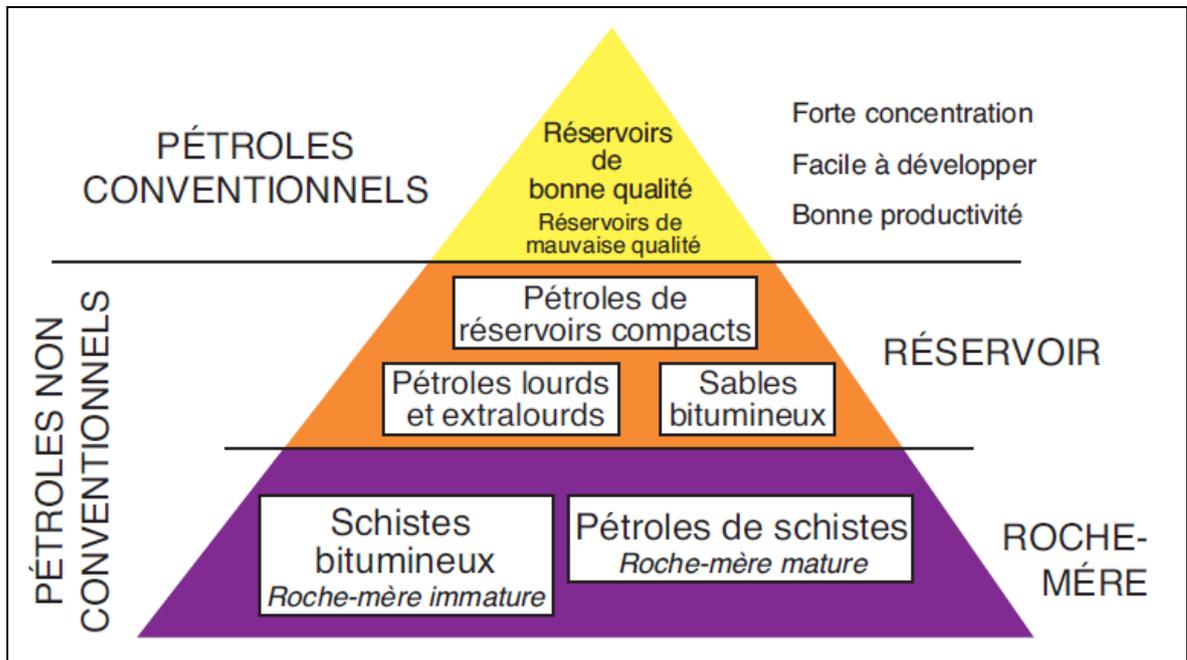


Fig. 03 : Le différent type d'hydrocarbures liquides non conventionnels (Rouaud, 2013).

a) Classification des hydrocarbures non conventionnels gazeux

Les ressources potentielles sont d'autant plus importantes que l'on descend dans le triangle.

Comme dans la représentation précédente, les hydrocarbures gazeux piégés dans des réservoirs de bonne qualité sont présentés à la pointe supérieure du triangle (en jaune), Au-dessous, et sans qu'il existe une limite franche, on trouve les gaz piégés dans des réservoirs compacts de très mauvaise qualité (en orange), En partie basse intermédiaire, on trouve :

- les gaz de houille (CBM, pour Coalbed Methane) : ces gaz (principalement du méthane), adsorbés sur le charbon, ont été parmi les premiers gaz non conventionnels mis en exploitation, aussi bien par forage que dans les mines ;
- les gaz de schiste (shale gas), prisonniers d'une roche-mère ; ils sont dits "secs" s'il ne s'agit que de méthane (gazeux sous conditions normales de température et de pression) ou "humides" s'ils contiennent aussi des gaz plus lourds, liquides sous conditions normales de température et de pression ;

Enfin, tout en bas du triangle, on trouve les hydrates de méthane, une combinaison solide de molécules d'eau et de méthane ("chlatrates"), stable dans la nature dans les zones arctiques ou dans les océans sous des tranches d'eau suffisamment importantes pour que des conditions de pression et de température favorables à leur formation puissent être rencontrées (Rouaud, 2013).

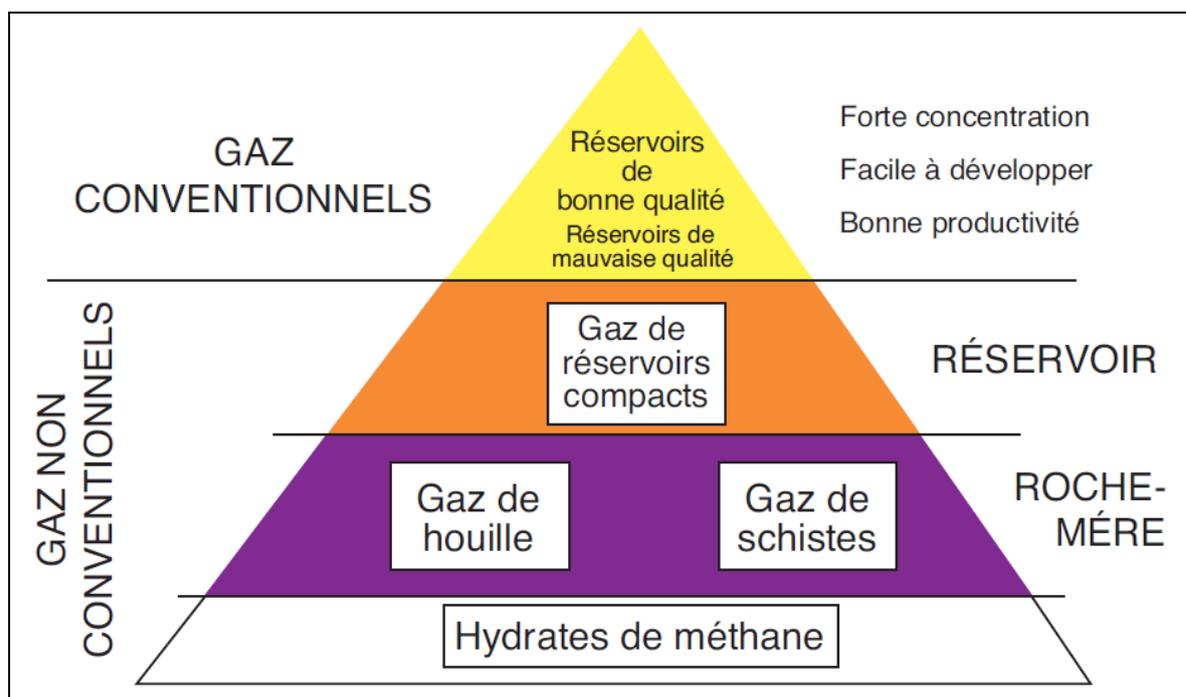


Fig. 04 : Les différents types d'hydrocarbures gazeux non conventionnel (Rouaud, 2013).

b) Les hydrocarbures de roche-mère

Comme présenté dans les paragraphes ci-dessus, les hydrocarbures de roches-mères sont de quatre natures différentes (fig. 05) :

- le gaz de houille, ou CBM,
- le gaz de schiste,
- le pétrole (ou huile) de schiste,
- les schistes bitumineux,

Le gaz de houille, comme on l'a déjà vu, est adsorbé sur du charbon, Ses caractéristiques sont développées dans le paragraphe,

Les trois autres, en revanche, forment en réalité un continuum : les limites entre schistes bitumineux, pétroles et gaz de schiste ne sont pas franches (fig.05), Tous trois sont générés dans des roches-mères, généralement des argiles ayant un fort contenu en matière organique, et ne se distinguent les uns des autres que par leur histoire géologique et l'état de maturité de la matière organique dont ils sont issus, cette maturité dépendant elle-même de la température atteinte au cours de l'enfouissement :

- si l'enfouissement a été insuffisant (moins de 2 000 m), la matière organique contenue dans les schistes n'est pas transformée en hydrocarbures, On parle alors de "**schistes bitumineux**" (oil shale), Pour produire des hydrocarbures à partir de ces roches, il est nécessaire de faire artificiellement ce que la nature n'a pas fait, On doit donc les chauffer afin de transformer la matière organique en hydrocarbures, Du point de vue du rendement énergétique et donc de l'intérêt économique, ces schistes bitumineux ne sont pas intéressants et n'ont été exploités que durant les conflits ou les embargos, En France, de telles couches ont été exploitées dans la région d'Autun,
- lorsque l'enfouissement a été un peu plus important (2 000 à 4 000 m), la matière organique a été transformée en hydrocarbures liquides ("Fenêtre à huile"), Ce pétrole a pu soit s'échapper des schistes pour former des gisements conventionnels, soit resté piégé dans les schistes, c'est alors le **pétrole de schiste**,
- lorsque l'enfouissement a été très important (supérieur à 4 000 m, "Fenêtre à gaz") la roche mère génère du gaz, La fraction de ce dernier qui reste piégée dans les schistes est le **gaz de schiste**,

Dans un même bassin sédimentaire, on peut ainsi avoir coexistence de gisements conventionnels, formés à partir des hydrocarbures expulsés, et d'hydrocarbures non conventionnels.

C'est d'ailleurs généralement le cas aux Etats-Unis, où les bassins sédimentaires qui produisent de grandes quantités d'hydrocarbures de roche-mère étaient déjà des bassins pétroliers ou gaziers, Cela pourrait être aussi le cas pour le bassin de Paris (Rouaud, 2013).

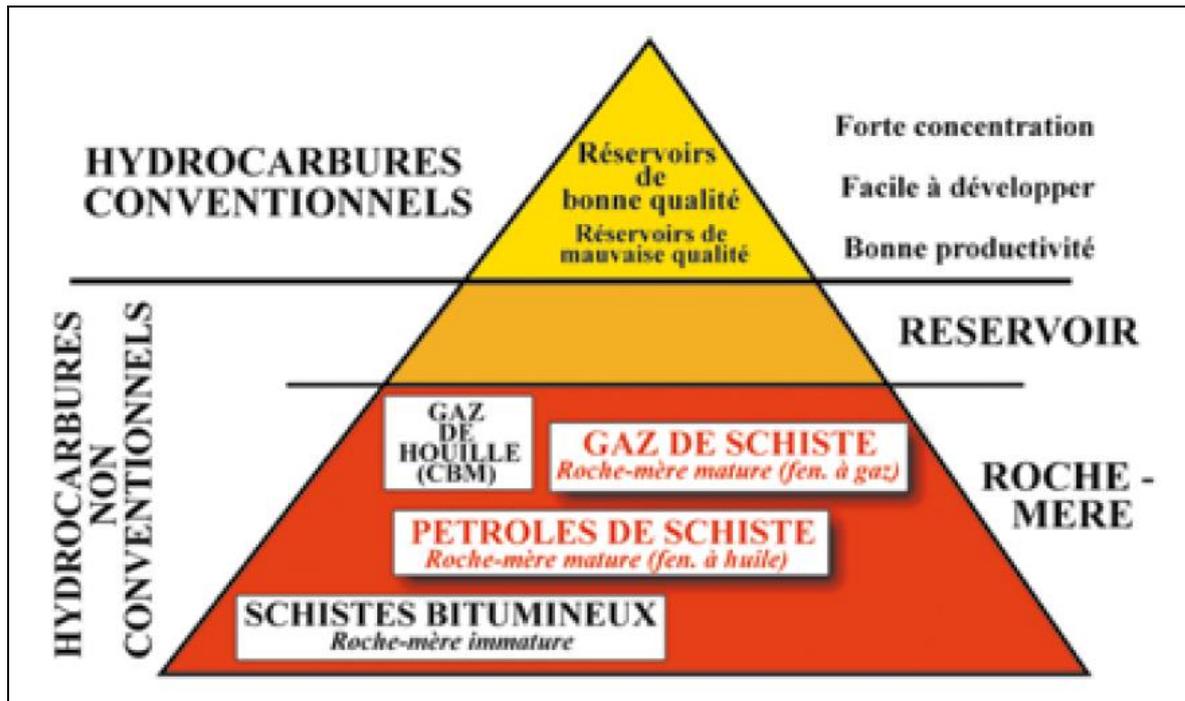


Fig. 05 : Les hydrocarbures de roche-mère (Rouaud, 2013).

L'accroissement de la demande énergétique mondiale repose sur deux tendances lourdes, l'augmentation de la population mondiale (1 % par an, nous serons 8,5 milliards en 2035) et la croissance économique (3 % par an à l'échelle mondiale). La part des hydrocarbures dans le mix énergétique mondial est appelée à diminuer mais elle devrait être encore de 50 % en 2035 (AIE). Les hydrocarbures non conventionnels sont appelés à fournir une part de ces hydrocarbures. Cette part ne pourra être importante que si l'empreinte environnementale de leur exploitation est réduite.

2. Etude d'un cas de schistes de Barnett (ou Barnett shale)

Introduction

A ce jour trois pays produisent de shale gas à l'échelle commerciale. Les Etats-Unis, Canada et la Chine. Aux Etats-Unis, le shale gas a été produit commercialement pour la

première fois en 1998, la production atteignait 20.6 milliards de mètre cubes par an soit 4% de la production totale de gaz naturel (Ibrahim and Abdelaziz, 2018).

Dans ce travail nous avons choisi le cas d'Etats-Unis d'un bassin au nord Texas de réservoir **schistes de Barnett** (ou Barnett shale) pour faire une **prédiction de carbone organique totale (COT) en utilisant la régression linéaire multiple**.

Le Barnett Shale est une région de l'Etat du Texas située dans la ville de Fort Worth près de Dallas. Elle tient ce nom de la formation géologique située dans le bassin Bend Arch-Fort Worth. Elle est devenue célèbre à la suite de la mise en valeur des grandes quantités de gaz non conventionnel ou shale gas qu'elle renferme. Elle est considérée comme la plus grande des régions de production du gas shale et sa superficie serait de 13000 km² (Ireland, 2012).

2.1. Contexte géologique et stratigraphique du bassin de Fort Worth

Barnett Shale est un dépôt du plateau marin du Mississippien dont l'épaisseur varie de 200 pieds (1 pied= 0.3048 m) dans la région sud-ouest à 1000 pieds dans le nord-est près de l'arche de Muenster. Et le Basin Fort Worth est situé dans le centre-nord du Texas, qui est délimité au nord, au nord-est et à l'est par des failles le soulèvement du sous-sol de l'Arche de la rivière Rouge, de l'arche de Muenster et du front structural d'Ouachita. La limite sud est définie par Llano Uplift (Abouelresh and Slatt, 2011).

Dans la région de Fort Worth, le Barnett est riche en matière organique (TOC 4,5%) et composé de roches à grains fins, non silicoclastiques, avec une perméabilité extrêmement faible. La matière organique dans le schiste pourrait atteindre 200 scf / tonne (Montgomery et al, 2005).

En outre, le schiste de Barnett est composé de deux intervalles de production notés comme Barnett supérieur et Barnett inférieur (Upper Barnett et Lower Barnett) qui se sont séparés par le calcaire de Forestburg dans cette région. Les données historiques indiquent que lorsque la production de Barnett inférieur et supérieur est mélangée, la contribution inférieure Barnett est 75-80% du total (Shelley et al, 2008).

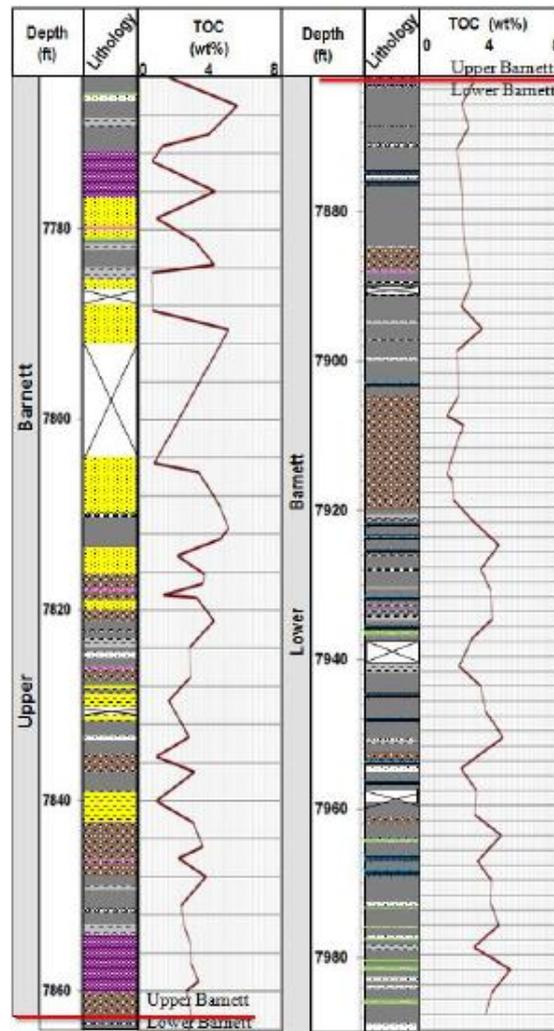


Fig. 06 : Log lithologique avec le COT% en poids correspondant, pour le schiste de Barnett (Abouelresh and Slatt, 2011).

La recherche stratigraphique indique que le schiste de Barnett dans la région de Fort Worth est entouré de carbonates serrés, le calcaire MarbleFall en haut et le calcaire Viola en bas (Janwadkar et al, 2006), agissant comme barrières de fracture pendant achèvement. Puisque le calcaire de Viola pince à l'ouest de la région de Fort Worth, la fracture hydraulique dans la partie inférieure de Barnett pourrait entrer dans le poreux Ellenberger, qui est une eau connue. Source, et conduire à la production d'eau élevée.

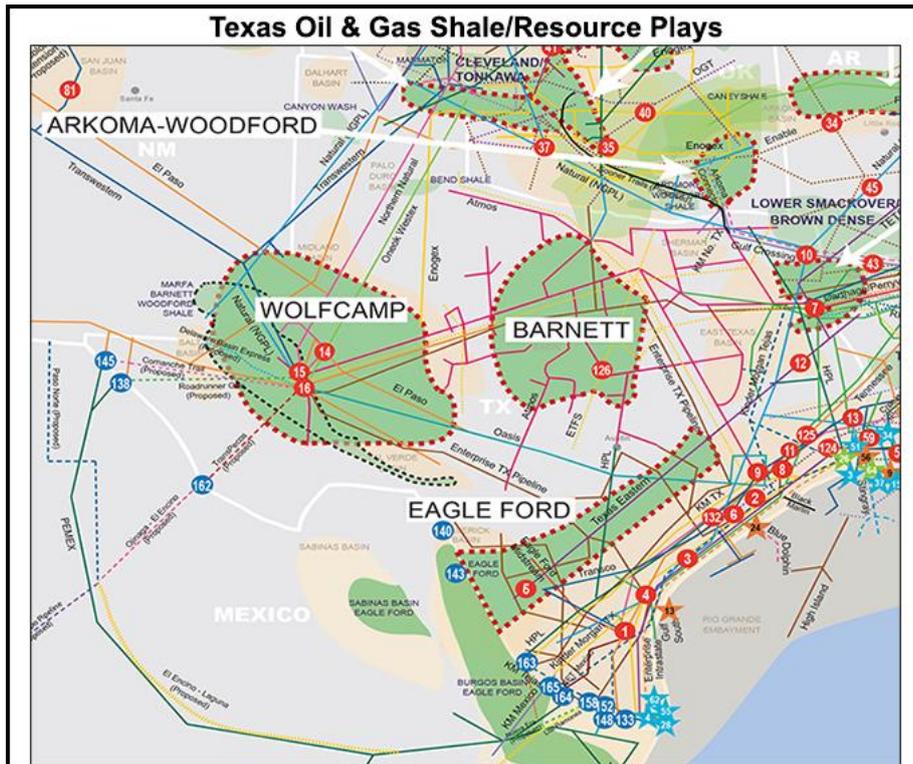


Fig. 07 : Carte de localisation du Barnett shale-bassin de Fort Worth (Natgaintel, 2016).

3. Généralités sur le carbone organique totale COT

Le carbone organique total (COT) est la mesure de la quantité de carbone disponible dans un composé organique et est souvent utilisé comme un facteur essentiel pour l'évaluation des ressources de schiste non conventionnelles. La quantité minimale de COT que doit contenir une roche pour être qualifié de roche mère à gaz est de 2 % (Ibrahim and Abdelaziz, 2018)

Tab. 01 : Teneur en COT définissant le potentiel pétrolier d'une roche mère.

Potentiel pétrolier	Teneur en carbone organique total
Pauvre	0.05 %
Faible	0.5 – 1%
Bon	1 – 2 %
Très bon	2 – 4 %
excellent	> 4 %

La teneur en COT diminue avec la maturation thermique, la nature même de la matière organique est aussi un facteur qui contrôle son potentiel pétrolier. La teneur élevée en

matière organique associée au shale est un bon indicateur du potentiel d'hydrocarbures des roches mères, elle est caractérisée par une résistivité élevée avec une certaine porosité apparente sur log analysais.



Fig. 08 : Roche mère de type shale (argileuse feuilletée), contenant de la matière organique (Ibrahim and Abdelaziz, 2018).

a) Origines de la matière organique

La matière organique est issue des êtres vivants : tous les végétaux et les animaux sont « construits » avec un ensemble de molécules principalement à base de carbone, d'azote et d'oxygène. A la surface de la Terre, la biomasse végétale représente plus de 90% de la biomasse totale : le rôle de la photosynthèse chez les végétaux est essentiel dans le cycle du carbone. C'est le phytoplancton qui est l'élément majeur de concentration du carbone dans les océans. Pollens, spores, algues, sont autant de constituants que l'on trouve dans les pétroles. De leur côté, les plantes continentales produisent aussi la cellulose et la lignine (TRANIER, 2006).

Issue de diverses molécules plus ou moins complexes (lipides, protides, glucides, mais aussi chlorophylle), la matière organique est très diversifiée et très fragile.

Pour aboutir à la formation d'hydrocarbures, elle devra donc trouver un milieu idéal, où la dégradation sera limitée, pour que l'accumulation et la sédimentation puissent avoir lieu.

b) Préservation de la matière organique

Le milieu doit d'abord contenir beaucoup de matière organique : plus il y aura de production, plus les chances de préservation seront grandes. Dans les océans, le plancton

ne vit que dans la première centaine de mètres de la tranche d'eau. Au delà, il n'y a plus assez de lumière pour que la photosynthèse ait lieu. Pour sédimenter, la matière organique doit atteindre le fond de la mer afin d'être enfouie et préservée (TRANIER, 2006).

Elle doit tout d'abord échapper à l'activité bactérienne, qui recycle la plupart des organismes dans la chaîne alimentaire. Or, les bactéries et les micro-organismes sont présents dans les eaux, mais aussi dans les premiers 10 cm de sédiments.

Seules quelques molécules miraculées atteignent ce stade dans un océan. Heureusement, certains milieux sont bien plus favorables à la préservation de la matière organique :

- les milieux anoxiques (très peu d'oxygène dissous), dans lesquels les tissus animaux et les produits organiques sont à l'abri des bactéries ;
- les milieux où le taux de sédimentation est fort : les sédiments s'accumulent très vite, protégeant ainsi la matière organique enfouie.

c) **Maturation de la matière organique**

L'évolution de la matière organique en kérogène puis en hydrocarbure, c'est-à-dire, sa maturation est le résultat d'un mécanisme physico-chimique complexe dont les principaux facteurs sont : l'enfouissement (pression) et la température (gradient géothermique). La maturité thermique mesure le degré d'exposition des formations la haute chaleur nécessaire pour la transformation de la matière organique en hydrocarbures. On distingue trois phases de maturation par lesquelles transite la matière organique depuis sa fossilisation jusqu'à sa dégradation finale. Les étapes successives sont : La diagenèse, La catagenèse et La métagenèse (Ibrahim and Abdelaziz, 2018).

➤ **La diagenèse**

Elle est caractérisée par une altération à basse température de la matière organique de 50 à 80°C (122 à 152°F) progressivement convertie en kérogène.

➤ **La catagenèse**

Se produit généralement lorsque l'enfouissement supplémentaire entraîne une augmentation de la pression, ce qui augmente la chaleur dans la plage d'environ 80 à 150°C (152 à 302°F), il y a craquage thermique. Les composés se scindent en molécules de plus en plus petites au fur et à mesure de l'enfouissement. La formation de pétrole proprement dit a lieu à ce moment, dans les conditions dites de « la fenêtre à huile ».

➤ **La métagenèse**

Est la dernière étape dans laquelle la chaleur et les changements chimiques entraînent presque la transformation du kérogène en carbone. Au cours de cette étape, le méthane tardif ou le gaz sec a évolué, avec des gaz non hydrocarbonés tels que CO₂, N₂ et H₂S. Plage de température d'environ 150°C à 200°C (302 à 392°F).

Conclusion

Dans ce chapitre on a parlé sur les gisements conventionnels et non conventionnels et on a montré que les seules différences entre eux est les conditions d'accumulations et les types de l'extraction.

On a parlé aussi sur

- Les différents types des hydrocarbures.
- Contexte géologique et stratigraphie du bassin de Fort Worth et sa richesse en matière organique. Et en fin on a fait une généralité sur le Carbone Organique Total (COT) et ces étapes de transformation en hydrocarbure (pétrole ou gaz).

Chapitre II

Matériel et Méthodes

Ce deuxième chapitre, *est composé de trois parties.*

En premier, nous allons définir les diagraphies et nous expliquerons en détails leur classification.

En *deuxième partie*, contient La préparation de la base de données qui est basée sur les données du COT mesurées en laboratoire et le calage avec les diagraphies.

En troisième partie, nous avons parlé de Méthodologie qui comporte les différentes modèles d'estimation du carbone organique total par la régression multiple.

Introduction

Dans ce chapitre ces matériel et méthodes on les utilisé pour élaborer le calcul du carbone organique total a partir des COT mesurés au laboratoire pour estimer les valeurs manquantes du COT (points non carotté) donc on a parlé sur

- Les diagraphies utilisées.
- Préparation de la base des données.
- La régression multiple.
- Feature engeneering.
- Validation externe.
- Les indices de performances.
- Et en fin les logiciels utilisés.

I. Les diagraphies utilisées (well logging)

C'est l'enregistrement en continu, le long du puits, des paramètres physiques du sous-sol (résistivité, radioactivité, vitesse acoustique, densité...) et leur interprétation en termes de caractéristiques géologiques (porosité, saturation en eau, argilosité, épaisseur, pendage, fracturation.

I.1. But des diagraphies

Les principaux objectifs des mesures diagraphiques sont les suivants :

- Estimer le contenu (les saturations) en eau et en hydrocarbures des réservoirs traversés ;
- Déterminer le pendage des couches ;
- Détermination des caractéristiques du puits de forage ;
- La comparaison des différents puits entre eux ;
- Repérer les couches productrices des couches isolantes (non-productrices) (O.SERRA, 1985).

I.2. Classification des diagraphies

Les diagraphies peuvent être subdivisées en quatre grands groupes :

- Diagraphies Electriques.
- Diagraphies Nucléaires.
- Diagraphies Acoustiques.
- Diagraphies Auxiliaires.
- Diagraphie de production.

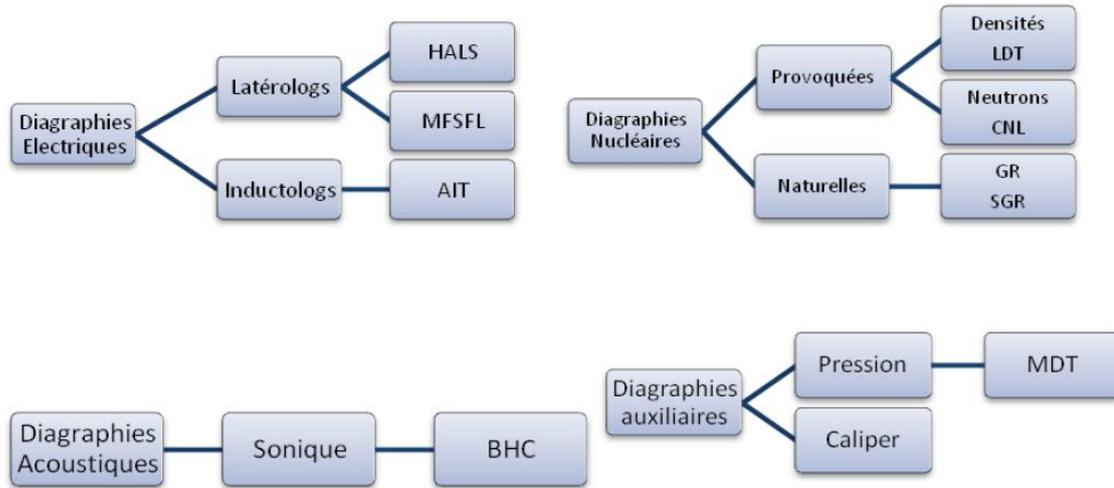


Fig. 09 : Schéma des différentes diagraphies (O.SERRA, 1985).

I.2.1. Diagraphies nucléaires

Les diagraphies nucléaires peuvent être enregistrées en trou ouvert ainsi qu'en trou tubé, puisque les rayons γ ou les neutrons traversent suivant leur énergie une épaisseur plus ou moins importante d'un matériau en fonction de sa nature (densité) (O.SERRA, 1985).

On regroupe sous le nom de diagraphies nucléaires :

- Le log de radioactivité gamma naturel, spectrométrie gamma.
- Le log gamma-gamma (densité).
- Etc.

a) Diagraphies gamma naturel

Les rayons gamma sont des salves d'ondes électromagnétiques de haute énergie qui sont émis spontanément par des éléments radioactifs. Presque tout rayonnement gamma rencontré sur la terre est émis par l'isotope radioactif de potassium et par les éléments radioactifs de la série d'uranium et du thorium.

Pour les formations sédimentaires on enregistre une radioactivité importante dans :

- Les formations argileuses qui renferment du potassium (illite spécialement) ;
- Les sels de potassium ;
- Les formations riches en matière organique peuvent concentrer l'uranium ;
- Les formations détritiques contenant des feldspaths (potassium) ou enrichie en minéraux lourds ;

➤ Principe

Le log GR est la mesure de la radioactivité naturelle de la formation. Dans les formations sédimentaires, le log reflète le contenu de cette formation en argile, parce que les éléments radioactifs sont généralement concentrés dans les argiles et schistes.

➤ Application des diagraphies gamma naturelles

- Détermination de la lithologie (argile, sels d'évaporite, minéraux lourds radioactifs) ;
- Etude minéralogique des argiles ;
- Estimation du pourcentage d'argile dans les formations gréseuses ;
- Les corrélations entre sondages et la détection de discordances ;
- Les recalages en profondeur ;

b) Diagraphies de densité (Gamma-Gamma)

C'est une diagraphie nucléaire provoquée qui utilise des sources radioactives pour l'émission des rayons gamma.

➤ Principe

On soumet la formation à un rayonnement α qui entre en collision avec la matière, et produit trois types d'interaction avec une perte d'énergie qui sont :

- Production Paire ;
- Effets Compton ;
- Effets photo électrique ;

➤ Applications des diagraphies Gamma-Gamma

- Calcul de la porosité ;
- Détermination de la lithologie ;
- Détermination de la zone à gaz (en la combinant aux diagraphies neutrons) ;
- Détection des minéraux lourds ;
- Détermination de la porosité utile en conjonction avec les diagraphies de porosité ;

I.2.2. Diagraphies électriques

La résistivité d'un milieu est la propriété physique qui détermine la capacité de ce milieu à laisser passer le courant électrique (Dr Mohamed Said Beghoul, 2013).

C'est la résistance ohmique d'un cylindre de longueur (L) et de section (S) : $\rho = R \cdot S / L$

- **R**- résistance électrique, ohm (Ω)
- **L**-longueur du cylindre

- S-section du cylindre
- ρ - résistivité, ohm ($\Omega.m$)



Fig. 10 : Une illustration montrant les différentes variables liées à la résistivité

➤ Principe de base de mesure

Le principe de base commun à toutes les diagraphies de résistivité consiste à envoyer un courant électrique (i) dans le sol et mesurer la différence de potentiel entre l'électrode de surface (référence) et celle(s) descendue(s) dans le trou de forage.

I.2.3. Diagraphie sonique (ondes acoustiques)

Le log sonique est basé sur l'étude de la propagation d'ondes acoustiques dans les formations. La mesure de la propagation de ces ondes et de leur atténuation apporte des renseignements sur les propriétés mécaniques des formations traversées (O.SERRA, 1985).

- La diagraphie sonique utilise principalement les ondes de compression (mouvement longitudinal des particules). Ces **ondes (P)** se propagent dans les solides, les liquides et les gaz, pour déterminer : de la porosité des roches, Détection des fractures, Corrélations lithologiques entre différents sondages et etc ;
- Les **ondes (S)** transversales ne se transmettent pas dans les liquides (Dr Mohamed Said Beghoul, 2013).

➤ Principe de mesure

Emission d'une onde de compression (P) et enregistrement du temps de transit (Δt) par des récepteurs (R1, R2) situés à une distance donnée de l'émetteur (3 à 5 pieds).

E- Emetteur

R1, R2- Récepteurs

P-Onde de compression (longitudinale)

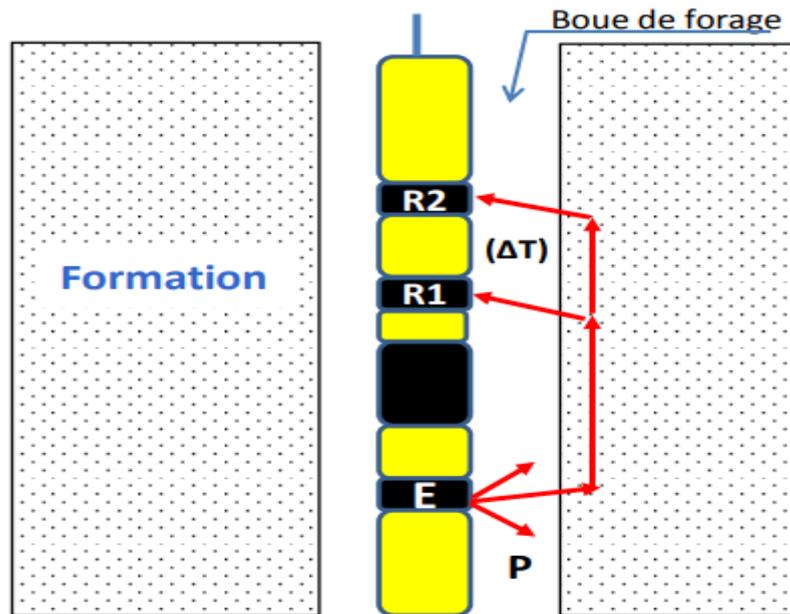


Fig. 11 : principe de mesure du Sonique .

II. Préparation de la base de données

La préparation de la base de données est basée sur les données du COT mesurées en laboratoire et le calage avec les diagraphies ci-dessus pour les intervalles carottés.

L'étude nécessitera des données numériques pour les diagraphies, mais, en raison de l'indisponibilité de ce dernier, les données graphiques trouvées dans l'article « New Technique to Determine the Total Organic Carbon Based on Well Logs Using Artificial Neural Network (White Box) (Abdulhamid et al., 2017) » ont été utilisées et converties en données numériques à l'aide du programme "GET DATA Graph Digitizer", puis placées dans le fichier "EXCEL" pour être prêt à l'emploi, comme le montre la figure ci-dessous.

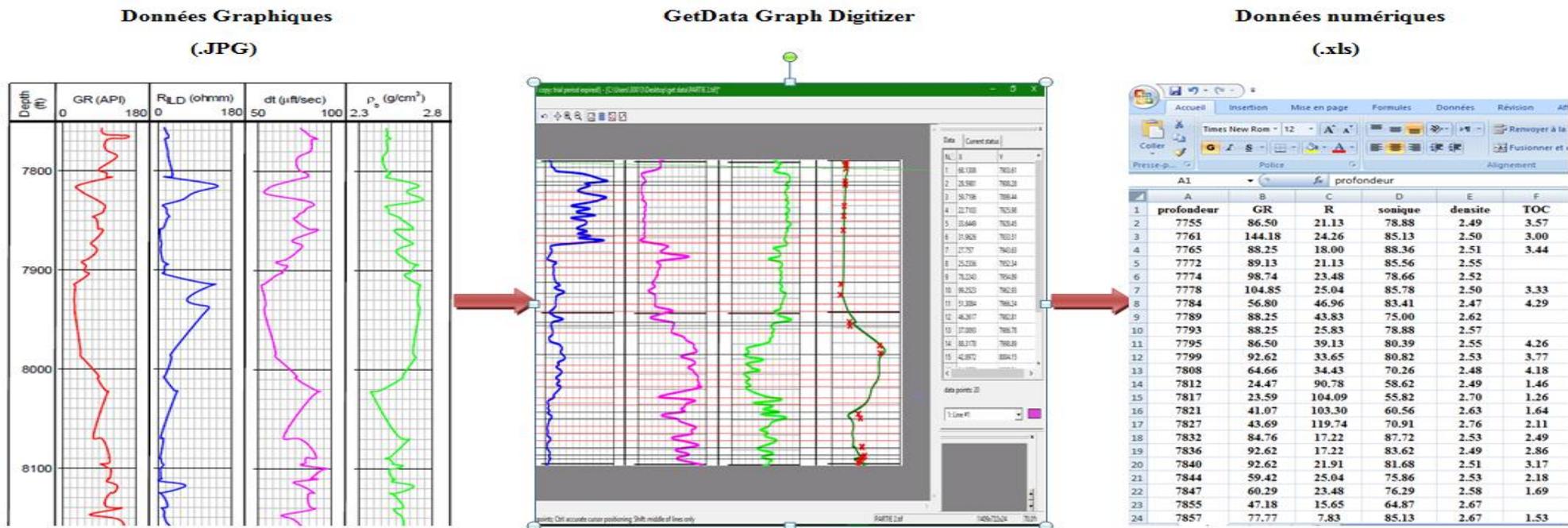


Fig. 12 : Conversion de données graphiques en données numériques.

III. Méthodologie

Le carbone organique total COT évalue la richesse de la roche mère en matière organique et sa prédisposition à générer des hydrocarbures. La teneur en COT permet de qualifier le potentiel roche mère. Il existe plusieurs méthodes de calculer le COT, notamment:

III. 1. Mesure directe

Se fait au laboratoire sur carotte, c'est un échantillon témoin d'une forme cylindrique extraits de réservoir à partir d'une opération s'appelle «le carottage », a partir de la carotte on peut déterminer le COT d'un réservoir après les analyses dans laboratoire, mais la récupération des carottes n'est pas toujours intégrale, à cause des problèmes techniques et géologiques rencontrés durant l'opération du carottage ainsi le cout et le temps de l'analyse au laboratoire.

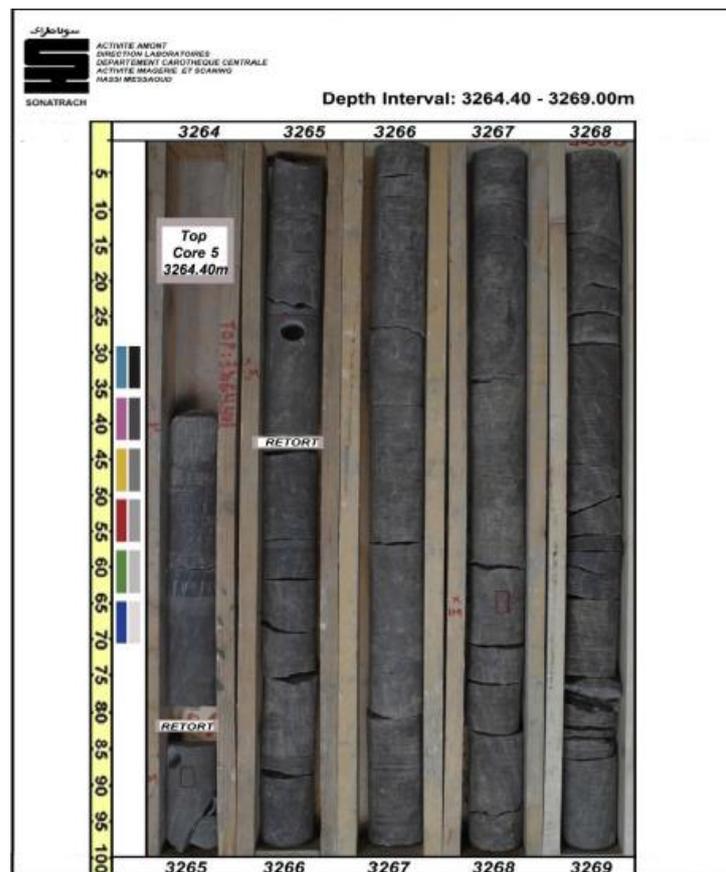


Fig. 13 : photo de carotte du puits (Sonatrach, 2010).

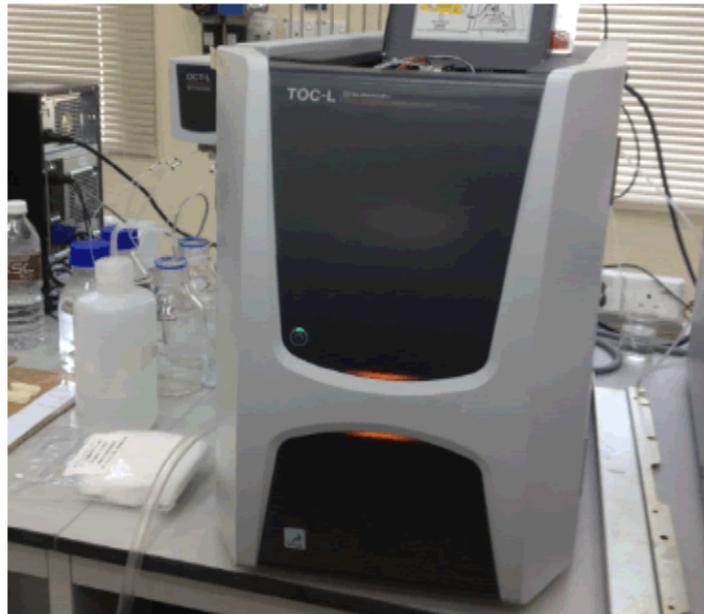


Fig. 14 : Dispositif de mesure du carbone organique total(COT) (Al-Selwi and Joshi, 2015)

III.2. Méthodes indirectes

Il y a différents modèles indirects permettent de calculer le COT :

III.2.1. La Régression Multiple

Le modèle de régression linéaire multiple est l'outil statistique le plus habituellement mis en œuvre pour l'étude de données multidimensionnelles. Cas particulier de modèle linéaire, il constitue la généralisation naturelle de la régression simple (WikiStat, no date).

Par exemple, une analyse de régression multiple peut révéler une relation positive entre la demande de lunettes de soleil et différents caractères démographiques (âge, salaire) des acheteurs de ce produit. La demande augmente et baisse avec les variations de ces caractéristiques.

➤ **Modèle théorique**

Étant donné un échantillon $(Y_i, X_{i1}, \dots, X_{ip})$ $i \in \{1, n\}$, on cherche à expliquer, avec le plus de précision possible, les valeurs prises par Y_i , dite variable endogène, à partir d'une série de variables explicatives X_{i1}, \dots, X_{ip} . Le modèle théorique, formulé en termes de variables aléatoires, prend la forme :

$$Y_i = a_0 + a_1X_{i1} + a_2X_{i2} + \dots + a_pX_{ip} + \varepsilon_i \quad i=1 \dots n$$

Où ε_i est l'erreur du modèle qui exprime, ou résume, l'information manquante dans l'explication linéaire des valeurs de Y_i à partir des X_{i1}, \dots, X_{ip} (problème de spécifications,

variables non prises en compte, etc.). Les coefficients a_0, a_1, \dots, a_p sont les paramètres à estimer.

➤ Estimation

Lorsqu'on dispose de n observations $(Y_i, X_{i1}, \dots, X_{ip}), i \in \{1, n\}$, qui sont des réalisations des variables aléatoires $(Y_i, X_{i1}, \dots, X_{ip})$, l'équation de régression s'écrit

$$Y_i = a_0 + a_1 X_{i1} + \dots + a_p X_{ip} + \varepsilon_i, \quad i=1, \dots, n$$

La problématique reste la même que pour la régression simple :

- estimer les paramètres a_0, a_1, \dots, a_p en exploitant les observations ;
- évaluer la précision de ces estimateurs ;
- mesurer le pouvoir explicatif du modèle ;
- évaluer l'influence des variables dans le modèle ;
- évaluer la qualité du modèle lors de la prédiction (intervalle de prédiction) ;
- détecter les observations qui peuvent influencer exagérément les résultats (points atypiques).

III.2.2. Feature Engineering (Transformation des données)

Le feature engineering renvoie à l'ensemble des interventions effectuées sur les données brutes avant qu'elles ne soient prises en compte par un algorithme apprenant (Alexandre Gerussi, 17/08/18).

Les transformations de données consistent donc en des manipulations permettant de révéler certaines caractéristiques des données originales qui ne peuvent pas ou pas facilement être observées. On peut procéder à la transformation de la distribution d'une variable de nos données pour la rendre plus facilement observable ou encore pour qu'elle corresponde aux hypothèses des tests statistiques que nous voulons appliquer. Une technique très fréquente consistera à appliquer une fonction mathématique sur notre variable. Par exemple, on pourra utiliser, plutôt qu'une variable originale X , son logarithme ou encore sa racine carrée.

III.2.2.1. Distribution normale et asymétrie des données

L'hypothèse la plus fréquente parmi tous les tests statistiques est celle qui requiert que la ou les variables analysées soient normalement distribuées. La distribution normale est aussi connue sous le nom de "courbe en cloche" (ou "bell curve" en anglais); On l'appelle ainsi parce que lorsque visualisées, les données prennent la forme d'une cloche de part et d'autre d'une valeur centrale. Par exemple, certaines mesures chez l'humain telles que la taille, le

poids, la durée de vie ou encore les résultats de tests de QI sont généralement normalement distribués autour d'une valeur moyenne de la population

Une distribution normale est répartie également de part et d'autre d'une valeur moyenne centrale. La courbe est symétrique par rapport à la moyenne. Lorsque la distribution des données est déséquilibrée soit à gauche ou soit à droite, on dira qu'elle est biaisée. Dans ces cas, il y a plus de données d'un côté du sommet de la courbe que de l'autre. Des données biaisées à droite ont une longue queue qui s'étend à droite de la moyenne de la distribution tandis que des données biaisées à gauche ont une longue queue à gauche. Lorsque des données sont très biaisées, il peut être difficile de visualiser les extrêmes. Dans ce cas, une transformation vous permettra soit de procéder à des tests statistiques ou encore à mieux visualiser la distribution en entier (F. DAUMAS, 1982).

III.2.2.2. Transformation logarithmique

La transformation logarithmique consiste tout simplement à appliquer la fonction log à chacune des valeurs de la variable à transformer. Vous utilisez ensuite dans votre analyse la nouvelle variable constituée des résultats de la fonction log au lieu des données originales. Cette transformation a un impact marqué sur la distribution d'une variable. Elle permet en général de rapprocher des valeurs extrêmes pour obtenir des graphes de distribution moins étendus. Les deux graphes ci-dessous montrent d'une part la variable population à laquelle on a appliqué la fonction logarithme naturel et, d'autre part, le même nuage de points que précédemment, mais auquel on a appliqué la fonction logarithme naturel sur les deux variables de population et superficie

III.2.2.3. Transformation Racine Carrée

La transformation racine carrée consiste à calculer la racine carrée de chacune des valeurs d'une variable d'un jeu de données et d'utiliser cette résultante dans notre analyse. Cette transformation a un impact plus modéré sur la distribution. Les deux graphes ci-dessous montrent la distribution de la population à laquelle on a appliqué une racine carrée, ainsi que le nuage de points, toujours de la population par rapport à la superficie, auquel on a appliqué la racine carrée aux deux variables (Alexandre Gerussi, 2018)

III.2.2.4. Transformation angulaire

Sur une distribution binomiale de probabilité p sur n observations $\arcsin(\sqrt{\hat{p}})$ on peut démontrer par la méthode du Delta que suit une loi normale d'espérance et de variance

$$\mu = \arcsin(\sqrt{p}) \quad \sigma^2 = \frac{1}{4n}$$

On peut donc assez simplement calculer un intervalle de confiance à 95 % par la formule (en appliquant une correction de continuité) :

$$\left[\Phi\left(\hat{p} - \frac{1}{2n}\right) - 1.96 \frac{1}{2\sqrt{n}} ; \Phi\left(\hat{p} + \frac{1}{2n}\right) + 1.96 \frac{1}{2\sqrt{n}} \right]$$

en posant :

$$\Phi(x) = \arcsin(\sqrt{x})$$

Ce calcul est surtout utile quand p/n se rapproche des extrêmes c'est à dire de 0 ou de 1

(Philippe MICHEL, no date).

III.2.3. D'Elimination des Outliers

Une valeur aberrante est une valeur extrême, anormalement différente de la distribution d'une variable, En d'autres termes, la valeur de cette observation diffère grandement des autres valeurs de la même variable,

Par exemple, si on prend la variable Salaire de la population française, l'ensemble des valeurs constituent une distribution de la variable Salaire, Imaginons qu'en général, pour cette distribution, les salaires varient entre 1600€ et 3000 €, Si on observe un individu avec un salaire à 10 000€, c'est une valeur extrême, On dira qu'il s'agit d'un outlier,

Dans le même sens, imaginons qu'une personne touche 200 € seulement, c'est un salaire anormalement bas, Il s'agit également d'une valeur aberrante (Younes Benzaki, no date).

III.2.4. La Validation externe

Il s'agit de modifier l'ordre des données d'une façon aléatoire n fois pour on obliger notre modèle de changer chaque fois entre les données de TRAINNAGES et les autres de VALIDATIONS.

IV. Les indices de performance

Une fois le modèle construit se pose alors la question s'il rend fidèlement compte de l'échantillon. Pour cela, la modélisation nécessite des critères de validation sur la base desquels, le concepteur juge les performances finales du modèle qui sera retenu. (Zaimeche et al., 2014).

❖ Le coefficient de détermination (R^2)

Celui-ci évalue le degré d'association entre deux variables et juge de la qualité de l'ajustement des valeurs observées et calculées :

$$R^2 = \left[\frac{\frac{1}{N} \sum_{i=1}^N (Y_{iobs} - \bar{Y}_{obs})(Y_{ical} - \bar{Y}_{cal})}{\sigma_{obs} \cdot \sigma_{cal}} \right]^2$$

Avec :

Y_{iobs} : Valeur observée.

Y_{ical} : Valeur calculée par le modèle.

\bar{Y}_{obs} : Moyenne de la variable observée.

\bar{Y}_{cal} : Moyenne de la variable calculée.

N : est le nombre de points.

Un coefficient de détermination proche de l'unité signifie que le modèle élaboré est très efficace.

❖ Erreur Quadratique Moyenne, Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^n (Y_{iobs} - Y_{isim})^2}$$

Quand le RMSE est proche de zéro signifie que le modèle est très bon.

❖ Erreur Absolue Moyenne, Mean Absolute Error (MAE)

$$MAE = \frac{1}{N} \sum_{i=1}^N |Y_{iobs} - Y_{isim}|$$

Lorsque le MAE est proche de zéro signifie que le modèle est très bon.

❖ **coefficient de corrélation (cc)**

$$CC = \frac{\frac{1}{N} \sum (Y_{iobs} - Y_{isim})(Y_{iobs} - Y_{isim})}{\sqrt{\frac{1}{N} \sum_{i=1}^n (Y_{iobs} - Y_{isim})^2} \sqrt{\frac{1}{N} \sum_{i=1}^n (Y_{iobs} - Y_{isim})^2}}$$

Un CC proche de 1 signifie que le modèle est très bon.

V. LOGICIELS UTILISES

31. Logiciel "GetData Graph Digitizer"

Get Data Digitizer est un programme utilisé pour numériser des tracés numérisés de données fonctionnelles. On trouve souvent des données présentées dans des rapports et des références sous forme de diagrammes en dispersion ou linéaires fonctionnels de type X-Y. Pour pouvoir utiliser ces données, elles doivent être numérisées. Ce programme vous permettra de prendre une image numérisée d'un tracé (au format GIF, JPEG ou PNG) et de numériser rapidement les valeurs du tracé en cliquant simplement avec la souris sur chaque point de données. Les numéros peuvent ensuite être enregistrés dans un fichier texte et utilisés où vous en avez besoin.

Get Data fonctionne avec des échelles à axes linéaires et logarithmiques. En plus de numériser des points hors des tracés de données, ce programme peut être utilisé pour numériser d'autres types de données numérisées (telles que des dessins à l'échelle ou des photos orthographiques).

32. Logiciel "EXCEL"

Ce logiciel qui est un tableur a été utilisé essentiellement pour la saisie et la gestion des fichiers des données. Les fichiers de données diagraphie de tout le réservoir ont été faits au préalable à l'aide de ce logiciel. En générale, il a été utilisé pour certains calculs (les indices de performance) et aussi pour qu'ils soient bien adapter à tout traitement statistique et se mettre en liaison avec Statistica.

33. Logiciel "STATISTICA"

STATISTICA est un portefeuille de logiciels qui offre la gestion des données, l'analyse des données, l'exploration de données et des solutions de visualisation de données. Il est un puissant programme avec des recettes intégrées, des assistants et smart et peut accéder aux données de presque tous les formats de fichiers.

Conclusion

On a résumé les étapes de la prédiction du carbone organique total (COT) dans cet organigramme.

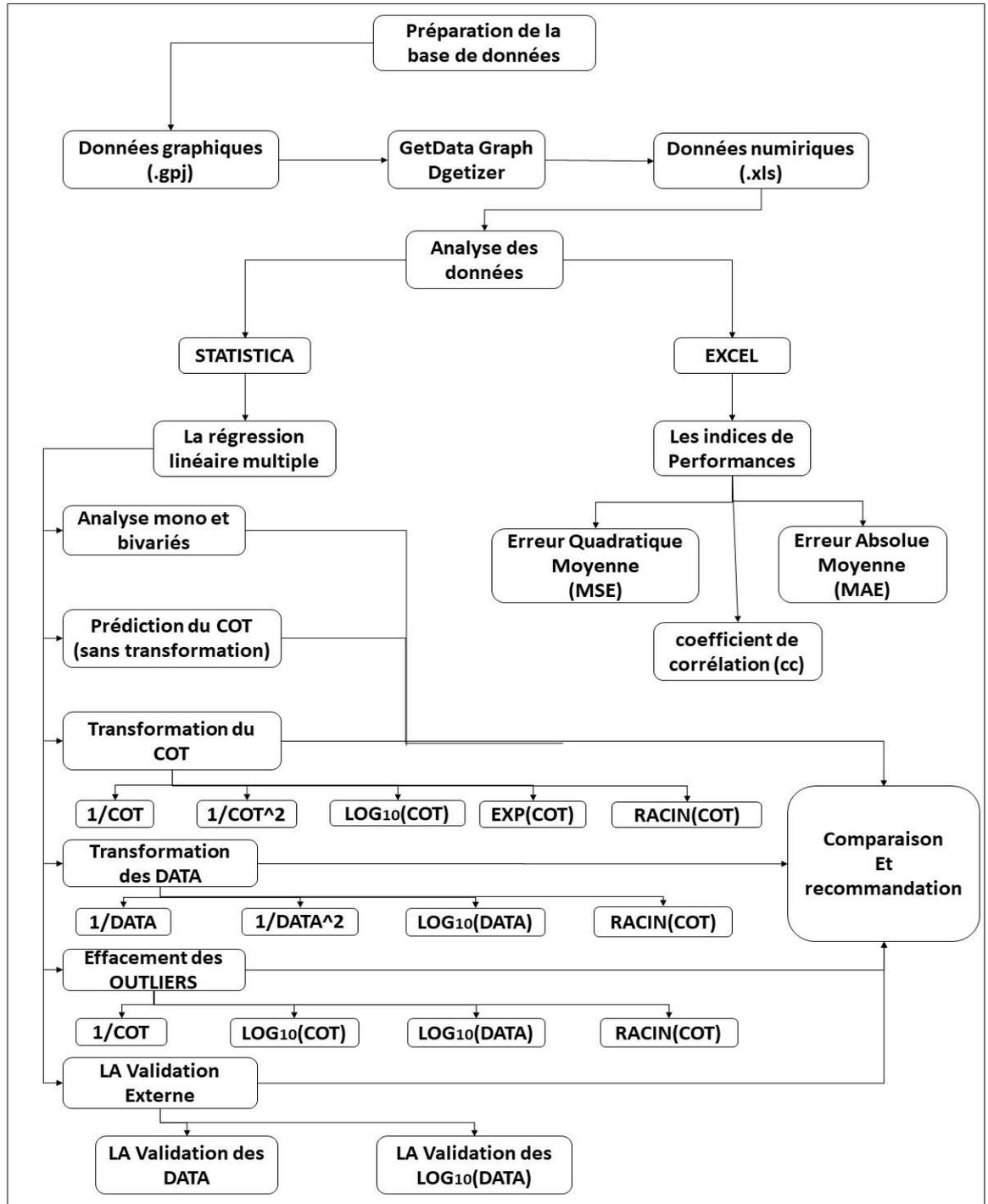


Fig. 15 : Organigramme explicatif de la prédiction du carbone organique total (COT).

Chapitre III

Résultats et Discussion

Le troisième chapitre présente les résultats obtenus par les différents modèles de la régression multiple et une comparaison qualitative et quantitative entre eux pour arriver au modèle le plus performant

Introduction

Ce chapitre contient un ensemble des résultats obtenus par plusieurs modèles de la régression multiple pour estimer le COT et une comparaison quantitative et qualitative pour atteindre le modèle le plus performante.

I. Analyse des données

On montre une base des données contient 103 points de mesures qui est représenté dans: Les enregistrements diagraphiques qui sont comme suit : (Gamma Ray, Résistivité, Densité et Sonique) et des valeurs du COT mesuré en laboratoire, mais malheureusement nous n'avons pas du COT dans tout le réservoir, ou il possède un intervalle de **52 points du COT mesurées en l'laboratoire** et un intervalle de **51 points du COT non mesurés**. une analyse de données ont été élaboré : Statistiques Descriptive ; mono et bivariés en utilisant le logiciel STATISTICA comme indiqué ci-dessous.

Tab.02: Tableau statistiques descriptive de COT mesuré et les enregistrements diagraphiques

	STATISTIQUES DESCRIPTIVES				
	Gamma Ray	Résistivité	Sonique	Densité	COT
Moyenne	86.5113	45.0233	75.0872	2.5771	2.3999
Médiane	89.1262	27.5000	77.80370	2.5697	2.2512
Mode	Multiple	7.5000	Multiple	Multiple	1.1775
Écart-type	36.2163	38.4965	11.2260	0.0844	1.1391
Variance	1311.626	1481.985	126.024	0.007	1.298
Minimum	23.5514	5.4782	52.3256	2.4000	0.7570
Maximum	168.1690	162.5000	95.1163	2.7565	4.3178
Somme	8910.669	4637.405	7733.983	265.445	124.798
Effectif	103	103	103	103	52

Tab. 03 : Matrice de corrélation de COT mesuré et les enregistrements diagraphiques

	MATRICE DE CORRELATION				
	Gamma Ray	Résistivité	Sonique	Densité	COT
Gamma Ray	1.00				
Résistivité	-0.35	1.00			
Sonique	0.76	-0.58	1.00		
Densité	-0.58	0.27	-0.64	1.00	
COT	0.65	-0.01	0.65	-0.76	1.00

Le calcul de la matrice de corrélation a été faite sur l'ensemble de 103 points de mesures qui convient au réservoir de gaz de schiste de Barnett (Tableau 03).

La matrice de corrélation a montré que le COT est corrélé positivement avec Gamma Ray ($R=0.65$) et le sonique ($R=0.65$) et négativement avec la densité ($R=-0.76$).

L'ensemble des corrélations calculées à partir de la formule ci-dessous sont significatives et sont marquées en rouge.

Formule :

$$R = 2/\sqrt{(n-1)}$$

1 Prédiction de COT par la méthode de Régression Multiple en différents modèles

1 Prédiction de COT avec les data brut (sans transformation)

➤ Comparaison qualitative

Le graph représente des résultats de l'Excel que est une comparaison entre le COT mesuré (en bleu) et le COT prédictible par la régression linéaire multiple (en orange).

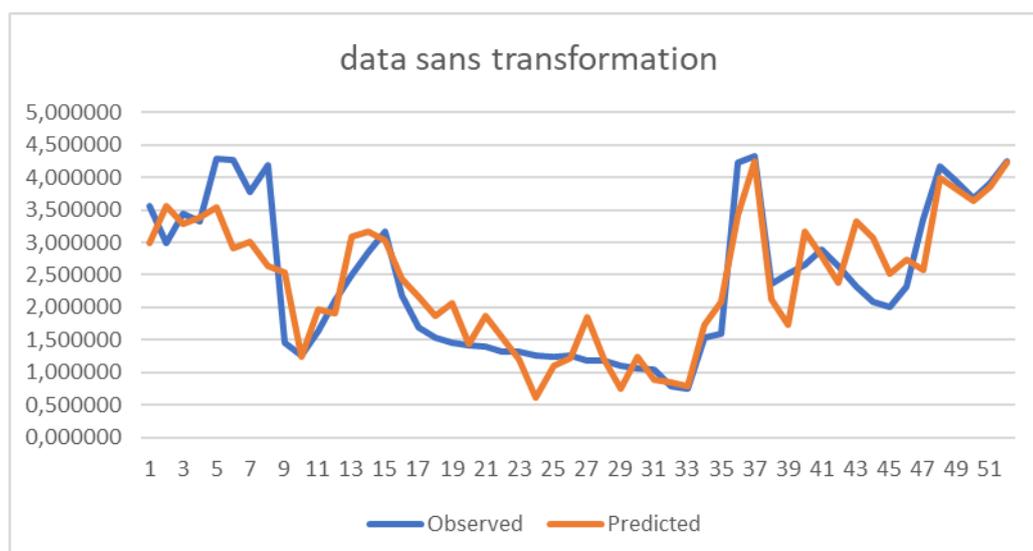


Fig. 16 : Comparaison graphique entre le COT mesuré et estimé par la méthode de RLM utilisant les data brut (sans transformation de variables).

➤ Comparaison quantitative

On a essayé avec cette méthode de trouver les valeurs du COT avec STATISTICA proche aux valeurs du COT mesurées utilisant les logs : Gama Ray, Résistivité, sonique et Densité. En utilisant la formule ci-dessous :

$$\text{COT} = 14,577 + 0.152 \text{ Gamma Ray} + 0.456 \text{ Résistivité} + 0.487 \text{ Sonique} - 0.482 \text{ Densité}$$

Selon le tableau statistique on distingue que :

- Les valeurs du COT estimées par cette méthode sont un peu faibles par rapport aux valeurs du COT mesurées au laboratoire.
- L'erreur quadratique moyenne (RMSE) et l'erreur absolue moyenne (MAE) sont un peu élevées (MSE =0.535, MAE =0.367), et un coefficient de corrélation presque approprié (CC =0.8801).

Tab. 04. Illustration de comparaison statistique entre le COT mesuré et estimé par la méthode de la RLM utilisant les data brut (sans transformation de variables).

	COT Mesurés	COT Simulés
obs 1	3,568180	2,988682
obs 2	2,998130	3,554185
obs 3	3,437000	3,280165
obs 4	3,331330	3,379260
obs 5	4,287140	3,545047
obs 6	4,259260	2,919147
obs 7	3,767190	3,011885
obs 8	4,180390	2,636374
obs 9	1,463300	2,541057
obs 10	1,255670	1,232551
obs 11	1,642710	1,959929
obs 12	2,107140	1,917689
CC	0,880197	
MSE	0,535417	
MAE	0,397401	

2 Prédiction de COT avec transformation de variables

Dans cette étape on va intégrer des formules mathématiques dans les valeurs des COT telle que :

- Inverse du COT
- inverse carré du COT
- Exponentielle du COT
- \log_{10} du COT
- Racine(COT)

1. Inverse du COT

Le graph représente des résultats de l'Excel qu'est une comparaison entre l'inverse du COT (1/COT) mesuré (en bleu) et l'inverse du COT prédictible par la régression linéaire multiple (en orange).

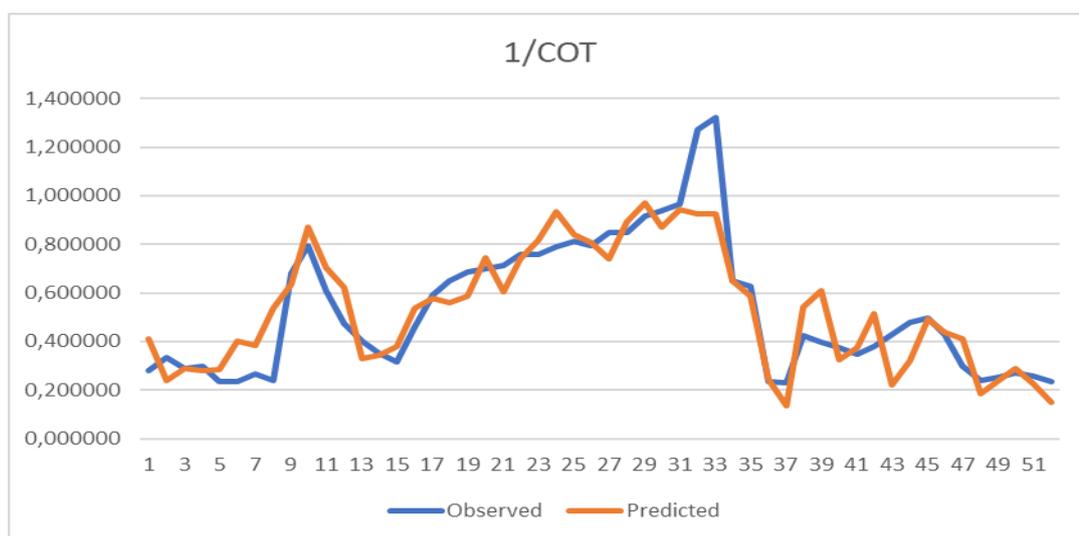


Fig. 17: Comparaison graphique entre le 1/COT mesuré et estimé par la méthode de RLM utilisant les data brut.

La formule de ce modèle est :

$$1/TOC = -0.168 \text{ Gamma Ray} -0.381 \text{ Résistivité} -0.642 \text{ Sonique} +0.333 \text{ Densité} -0.855$$

Tab. 05. Illustration de comparaison statistique entre le 1/COT mesuré et estimé par la méthode de la RLM utilisant les data brut.

	COT Mesurés	COT Simulés
obs 1	0,280255	0,410260
obs 2	0,333541	0,239485
obs 3	0,290951	0,289722
obs 4	0,300180	0,281938

obs 5	0,233256	0,285369
obs 6	0,234783	0,401829
obs 7	0,265450	0,382768
obs 8	0,239212	0,539123
obs 9	0,683387	0,637233
obs 10	0,796388	0,870024
obs 11	0,608750	0,704985
CC	0,895434	
MSE	0,120804	
MAE	0,086990	

2. LOG10(COT)

Le graph représente des résultats de l'Excel qu'est une comparaison entre le LOG₁₀(COT) mesuré et le LOG₁₀(COT) prédictible par la régression linéaire multiple.

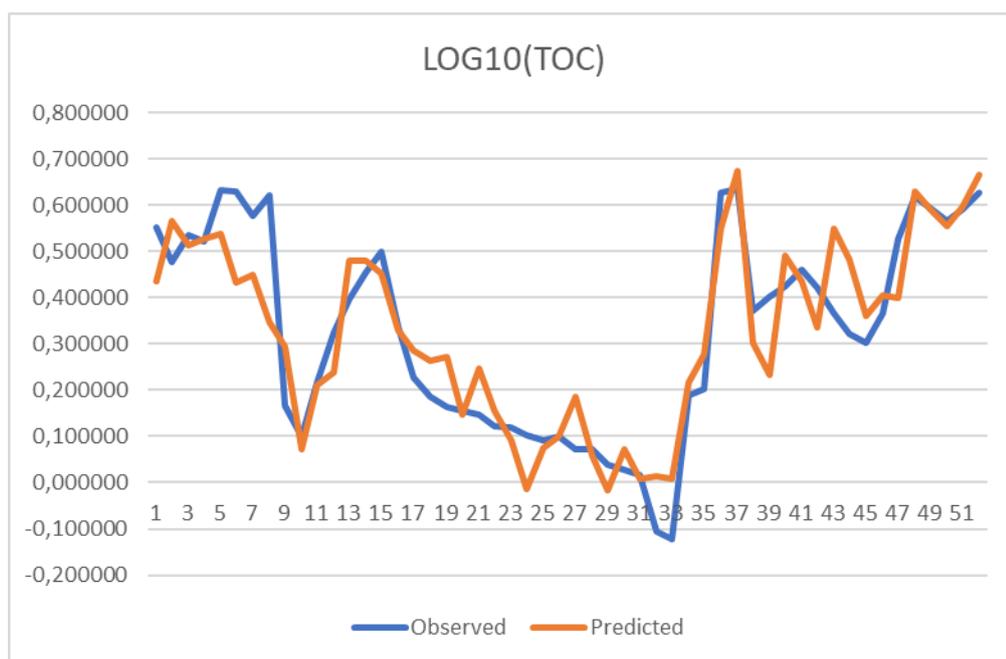


Fig.18 : Comparaison graphique entre le LOG₁₀(COT) mesuré et estimé par la méthode de RLM utilisant les data brut.

La formule de ce modèle est :

$$\text{LOG}_{10}(\text{TOC}) = +0.179 \text{ GR} + 0.423 \text{ Résistivité} + 0.570 \text{ Sonique} - 0.414 \text{ Densité} + 2.083$$

Tab. 06. Illustration de comparaison statistique entre LOG₁₀(COT) mesuré et estimé par la méthode de la RLM utilisant les data brut

	COT Mesurés	COT Simulés
obs 1	0,552447	0,436418
obs 2	0,476850	0,566623
obs 3	0,536180	0,513743
obs 4	0,522618	0,528295
obs 5	0,632168	0,538240
obs 6	0,629334	0,432718
obs 7	0,576018	0,450213
obs 8	0,621217	0,347588
obs 9	0,165333	0,294227
obs 10	0,098876	0,071188
obs 11	0,215561	0,211090
obs 12	0,323693	0,239544
CC	0,907702	
MSE	0,091298	
MAE	0,070049	

3. EXP(COT)

Le graph représente des résultats de l'Excel qu'est une comparaison entre le EXP(COT) mesuré et le EXP(COT) prédictible par la régression linéaire multiple.

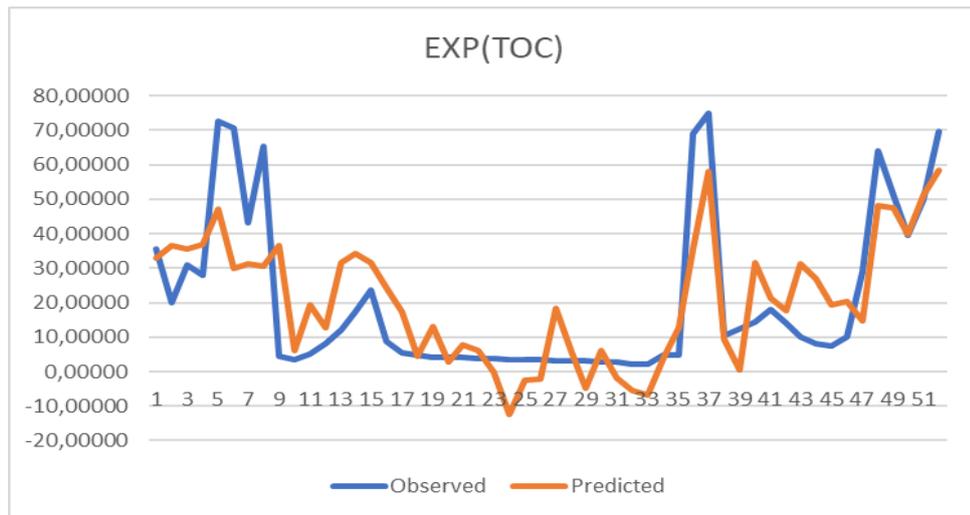


Fig. 19: Comparaison graphique entre le EXP(COT) mesuré et estimé par la méthode de RLM utilisant les data brut.

La formule de ce modèle est :

$$\text{EXP(TOC)} = -0.013 \text{ GR} + 0.490 \text{ Résistivité} + 0.368 \text{ Sonique} - 0.573 \text{ Densité} + 359.244$$

Tab. 07. Illustration de comparaison statistique entre EXP(COT) mesuré et estimé par la méthode de la RLM utilisant les data brut

	COT Mesurés	COT Simulés
obs 1	35,45201	32,8339
obs 2	20,04801	36,4352
obs 3	31,09354	35,5352
obs 4	27,97552	36,7981
obs 5	72,75808	47,1888
obs 6	70,75761	29,9338
obs 7	43,25834	31,3876
obs 8	65,39135	30,4751
obs 9	4,32019	36,6527
obs 10	3,51019	5,9704
obs 11	5,16916	19,2886
obs 12	8,22469	12,8341
CC	0,77523	
MSE	14,50069	
MAE	11,04469	

34. Racine(COT)

Le graph représente des résultats de l'Excel que est une comparaison entre la Racine(COT) mesuré et la Racine(COT) prédictible par la régression linéaire multiple.

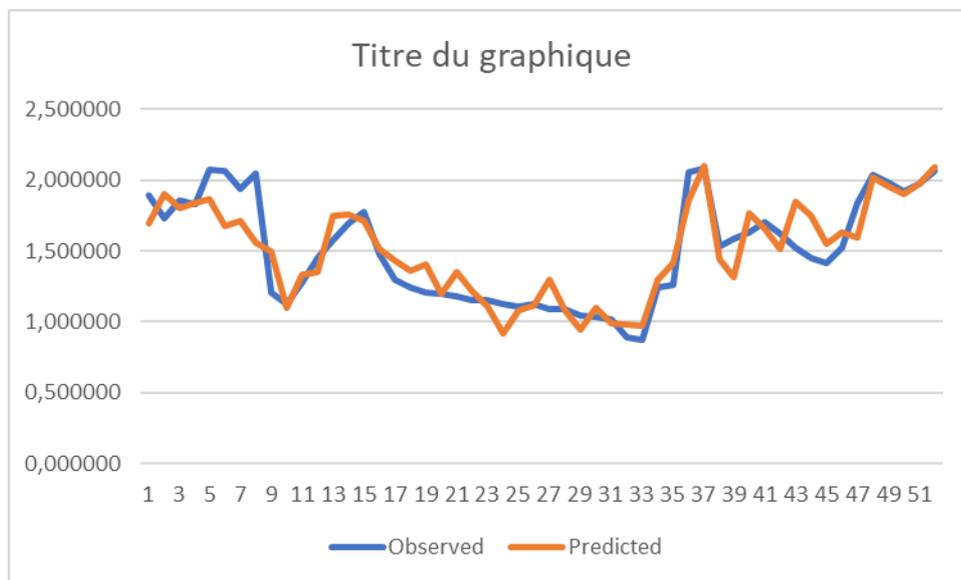


Fig. 20 : Comparaison graphique entre la Racine(COT) mesuré et estimé par la méthode de RLM utilisant les data brut (sans transformation de variables).

La formule de ce modèle est :

$$\text{RACINE(TOC)} = +0.169 \text{ GR} + 0.441 \text{ Résistivité} + 0.528 \text{ Sonique} - 0.450 \text{ Densité} + 4,993$$

Tab. 08. Illustration de comparaison statistique entre la Racine(COT) mesuré et estimé par la méthode de la RLM utilisant les data brut

	COT Mesurés	COT Simulés
obs 1	1,888963	1,693045
obs 2	1,731511	1,897318
obs 3	1,853915	1,805422
obs 4	1,825193	1,834776
obs 5	2,070541	1,868866
obs 6	2,063797	1,678211
obs 7	1,940925	1,708439
obs 8	2,044600	1,560106
obs 9	1,209669	1,498736
obs 10	1,120567	1,095389

obs 11	1,281682	1,333269
obs 12	1,451599	1,349313
CC	0,897069	
MSE	0,162405	
MAE	0,122891	

4. 1/TOC^2

Le graph représente des résultats de l'Excel qu'est une comparaison entre le 1/TOC² mesuré et le 1/TOC² prédictible par la régression linéaire multiple.

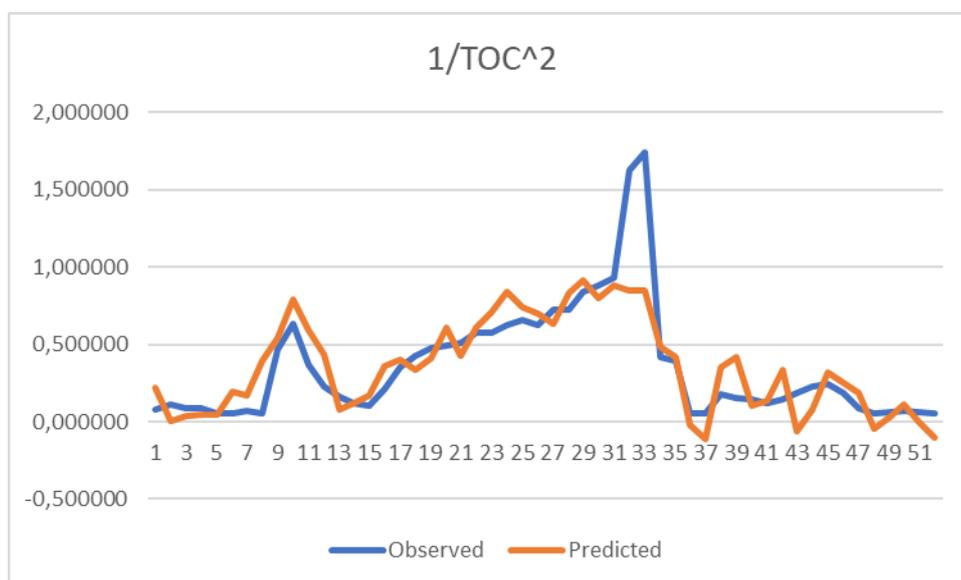


Fig. 21 : Comparaison graphique entre le 1/TOC² mesuré et estimé par la méthode de RLM utilisant les data brut.

La formule de ce modèle est :

$$1/(TOC)^2 = -0.124 \text{ Gamma Ray} -0.339 \text{ Résistivité} -0.668 \text{ Sonique} +0.253 \text{ Densité} -0.589$$

Tab. 09. Illustration de comparaison statistique entre 1/TOC² mesuré et estimé par la méthode de la RLM utilisant les data brut

	COT Mesurés	COT Simulés
obs 1	0,078543	0,217517
obs 2	0,111250	0,005320
obs 3	0,084653	0,037964
obs 4	0,090108	0,043078

obs 5	0,054408	0,049245
obs 6	0,055123	0,191372
obs 7	0,070464	0,171997
obs 8	0,057222	0,394928
obs 9	0,467018	0,538734
obs 10	0,634233	0,787682
obs 11	0,370577	0,590975
obs 12	0,225223	0,437244
CC	0,825671	
MSE	0,206045	
MAE	0,133843	

3 Prédiction de COT transformé avec transformation des DATA

Dans ce cas on a intégré les transformations dans tous les DATA pour but de voir est ce qu'il y a des changements les formules utilisées sont :

- Inverse des data
- Inverse² des data
- LOG₁₀(data)
- Racine data

1. Inverse des data

Le graph représente des résultats de l'Excel qu'est une comparaison entre l'inverse du COT mesuré et l'inverse du COT prédictible par la régression linéaire multiple. Utilisant dans ce cas l'inverse des DATA

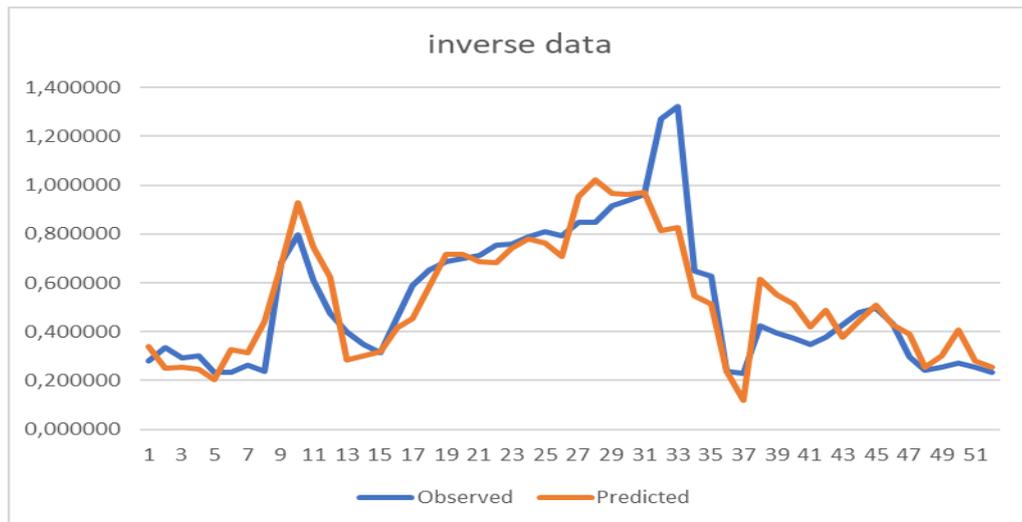


Fig. 22 : Comparaison graphique entre le 1/COT mesuré et estimé par la méthode de RLM utilisant l'inverse des DATA

La formule de ce modèle est

$$1/COT = 0.0054 * 1/GR + 0.3492 * 1/ R + 0.7429 * 1/ S - 0.2718 * 1/ D + 1.4808$$

Tab. 10. Illustration de comparaison statistique entre 1/COT mesuré et estimé par la méthode de la RLM utilisant l'inverse des DATA

	COT Mesurés	COT Simulés
obs 1	0,280255	0,339233
obs 2	0,333541	0,251971
obs 3	0,290951	0,255185
obs 4	0,300180	0,245790
obs 5	0,233256	0,205666
obs 6	0,234783	0,328372
obs 7	0,265450	0,314589
obs 8	0,239212	0,441620
obs 9	0,683387	0,673608
obs 10	0,796388	0,929290
obs 11	0,608750	0,746617
obs 12	0,474577	0,624653
CC	0,881963	
MSE	0,127893	
MAE	0,086395	

2. Inverse^2 des data

Le graph représente des résultats de l'Excel qu'est une comparaison entre l'inverse carré du COT mesuré et l'inverse carré du COT prédictible par la régression linéaire multiple. Utilisant dans ce cas l'inverse carré des DATA.

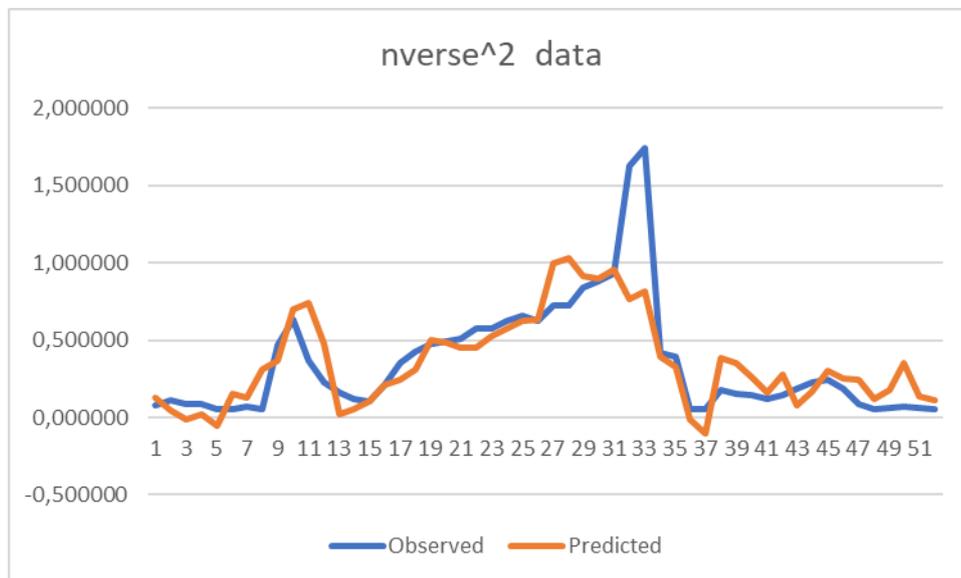


Fig. 23 : Comparaison graphique entre le 1/COT^2 mesuré et estimé par la méthode de RLM utilisant l'inverse^2 des data.

La formule de ce modèle est :

$$1/COT^2 = -0.371*1/GR^2 + 0.178*1/R^2 + 0.972*1/S2 - 0.238*1/D^2 + 0.750$$

Tab. 11. Illustration de comparaison statistique entre 1/COT^2 mesuré et estimé par la méthode de la RLM utilisant l'inverse^2 des data.

	COT Mesurés	COT Simulés
obs 1	0,078543	0,132271
obs 2	0,111250	0,041955
obs 3	0,084653	-0,009435
obs 4	0,090108	0,022408
obs 5	0,054408	-0,050661
obs 6	0,055123	0,153413
obs 7	0,070464	0,130304
obs 8	0,057222	0,313548
obs 9	0,467018	0,364035
obs 10	0,634233	0,700782

obs 11	0,370577	0,741998
obs 12	0,225223	0,475362
CC	0,802008	
MSE	0,218157	
MAE	0,132337	

3. LOG₁₀(data)

Le graph représente des résultats de l'Excel que est une comparaison entre le LOG₁₀(COT) mesuré et le LOG₁₀(COT) prédictible par la régression linéaire multiple. Utilisant LOG₁₀(data).

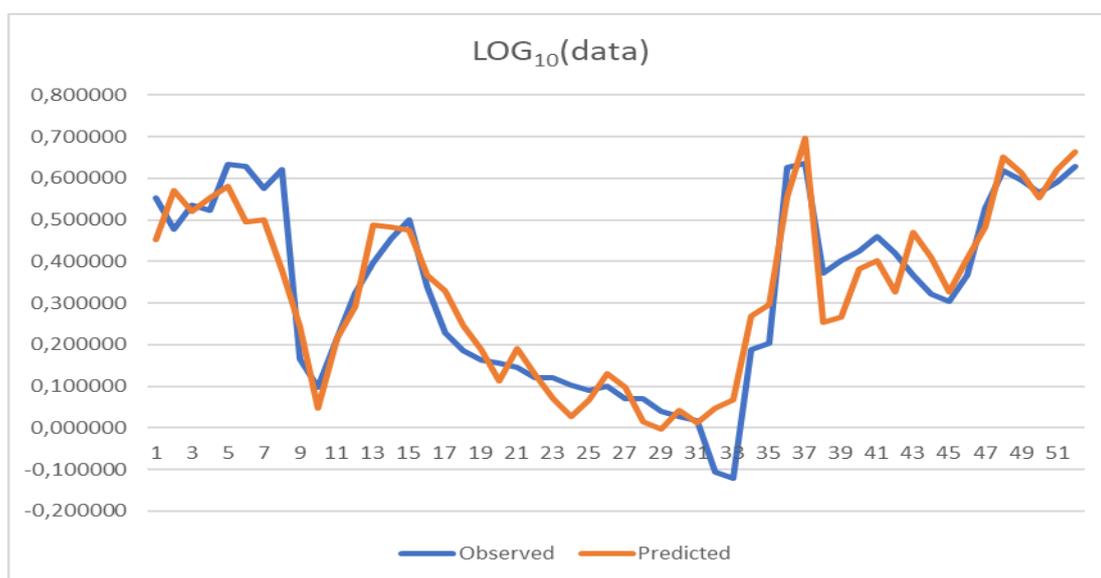


Fig. 24 : Comparaison graphique entre le LOG₁₀(COT) mesuré et estimé par la méthode de RLM utilisant LOG₁₀(data).

La formule de ce modèle et :

$$\text{LOG}_{10}(\text{COT}) = \mathbf{0.189} * \text{LOG}_{10}(\text{GR}) + \mathbf{0.520} * \text{LOG}_{10}(\text{R}) + \mathbf{0.659} * \text{LOG}_{10}(\text{S}) - \mathbf{0.304} * \text{LOG}_{10}(\text{D}) + \mathbf{1.4808}$$

Tab. 12. Illustration de comparaison statistique entre LOG₁₀(COT) mesuré et estimé par la méthode de la RLM utilisant LOG₁₀(data).

	COT Mesurés	COT Simulés
obs 1	0,552447	0,451324
obs 2	0,476850	0,570387
obs 3	0,536180	0,520163

obs 4	0,522618	0,552532
obs 5	0,632168	0,581799
obs 6	0,629334	0,494345
obs 7	0,576018	0,500904
obs 8	0,621217	0,380195
obs 9	0,165333	0,242570
obs 10	0,098876	0,047590
obs 11	0,215561	0,215476
obs 12	0,323693	0,292134
CC	0,933868	
MSE	0,077809	
MAE	0,061646	

4. Racine data

Le graph représente des résultats de l'Excel que est une comparaison entre la Racine(COT) mesuré et la Racine(COT) prédictible par la régression linéaire multiple. Utilisant la Racine de data

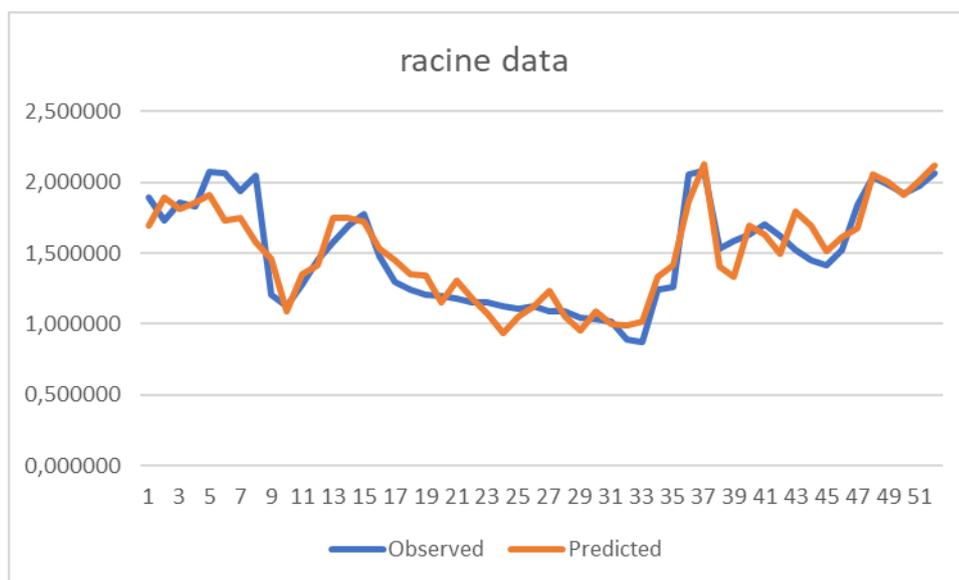


Fig. 25 : Comparaison graphique entre la Racine(COT) mesuré et estimé par la méthode de RLM utilisant la Racine de data.

La formule de ce modèle est :

$$\text{RACINE (COT)} = 0.175 * \text{RACINE(GR)} + 0.511 * \text{RACINE(R)} + 0.595 * \text{RACINE(S)} - 0.389 * \text{RACINE(D)} + 6.802$$

Tab. 13. Illustration de comparaison statistique entre la Racine(COT) mesuré et estimé par la méthode de la RLM utilisant la Racine de data

	COT Mesurés	COT Simulés
obs 1	1,888963	1,693235
obs 2	1,731511	1,896897
obs 3	1,853915	1,807169
obs 4	1,825193	1,852181
obs 5	2,070541	1,910069
obs 6	2,063797	1,730749
obs 7	1,940925	1,748629
obs 8	2,044600	1,573477
obs 9	1,209669	1,461185
obs 10	1,120567	1,089533
obs 11	1,281682	1,347303
obs 12	1,451599	1,417887
CC	0,917500	
MSE	0,146177	
MAE	0,114171	

Tab. 14. comparaison statistique entre les différents modèles de la RLM

model	TOC	1/TOC	LOG10(TOC)	EXP(TOC)	Racine TOC	1/TOC^2	inverse data	inverse^2 data	LOG10(data)	racine data
CC	0,880197	0,895434	0,907702	0,77523	0,897069	0,825671	0,881963	0,802008	0,933868	0,917500
MSE	0,535417	0,120804	0,091298	14,50069	0,162405	0,206045	0,127893	0,218157	0,077809	0,146177
MAE	0,397401	0,086990	0,070049	11,04469	0,122891	0,133843	0,086395	0,132337	0,061646	0,114171

Le tableau montre les indices de performance du différent modèle et différente transformation en remarque que y a d'améliorations dans certains modèles par rapport au COT mesuré sans transformation.

On prend 04 meilleur résultats pour supprimer sont OUTLIERS. Et discuté les nouveaux résultats

4 D'Elimination des OUTLIERS

On a Enlevé les OUTLIERS pour 4 meilleurs modèles que sont : 1/TOC, LOG10(TOC), LOG10(data) et racine data pour but d'amélioré encor plus les résultats.

1. Inverse du COT

On a enlevé les observation 08 ;32 ;33 et 43

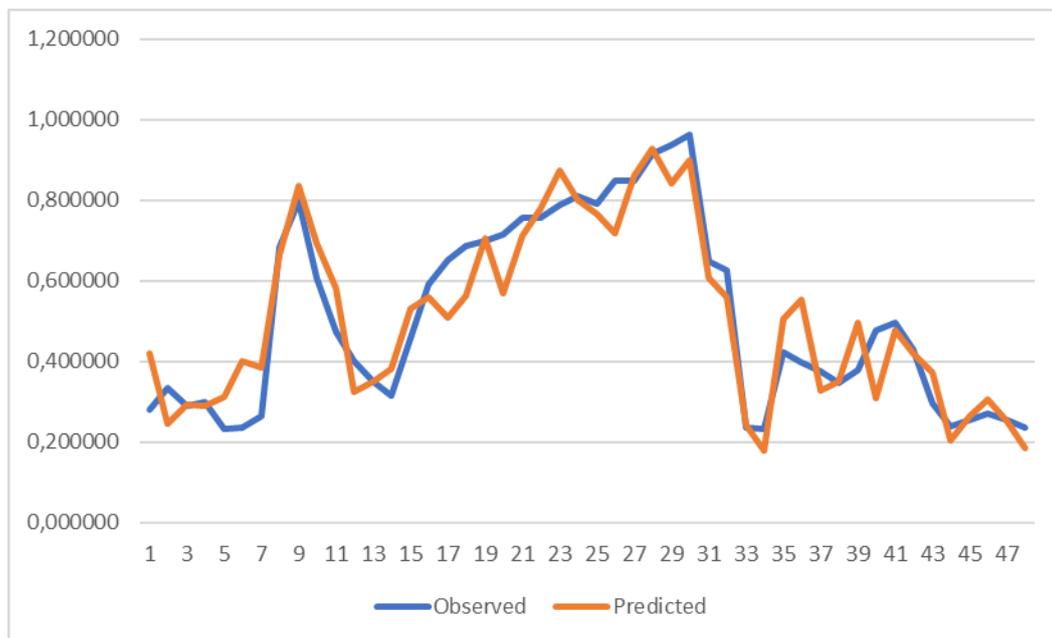


Fig. 26 : Comparaison graphique entre le 1/COT mesuré et estimé par la méthode de RLM utilisant les data brut après la d'élimination des outliers.

la formule de calcul de ce modèle est :

$$1/COT = -0.21 \text{ Gamma Ray } -0.41 \text{ Résistivité } -0.72 \text{ Sonique } +0.28 \text{ Densité } -0.0975$$

Tab. 15. Illustration de comparaison statistique entre 1/COT mesuré et estimé par la méthode de la RLM utilisant les data brut après la d'élimination des outliers

1/COT	COT Mesurés	COT Simulés
obs 1	0,280255	0,419872
obs 2	0,333541	0,244802
obs 3	0,290951	0,294005
obs 4	0,300180	0,289177
obs 5	0,233256	0,313270
obs 6	0,234783	0,399567
obs 7	0,265450	0,383846

obs 9	0,683387	0,665111
obs 10	0,796388	0,837584
obs 11	0,608750	0,692370
obs 12	0,474577	0,581852
obs 13	0,400924	0,325869
CC	0,937386	
MSE	0,080024	
MAE	0,062791	

2. LOG₁₀ (COT)

On a enlevé les observations 06 ;08 ;43et 44

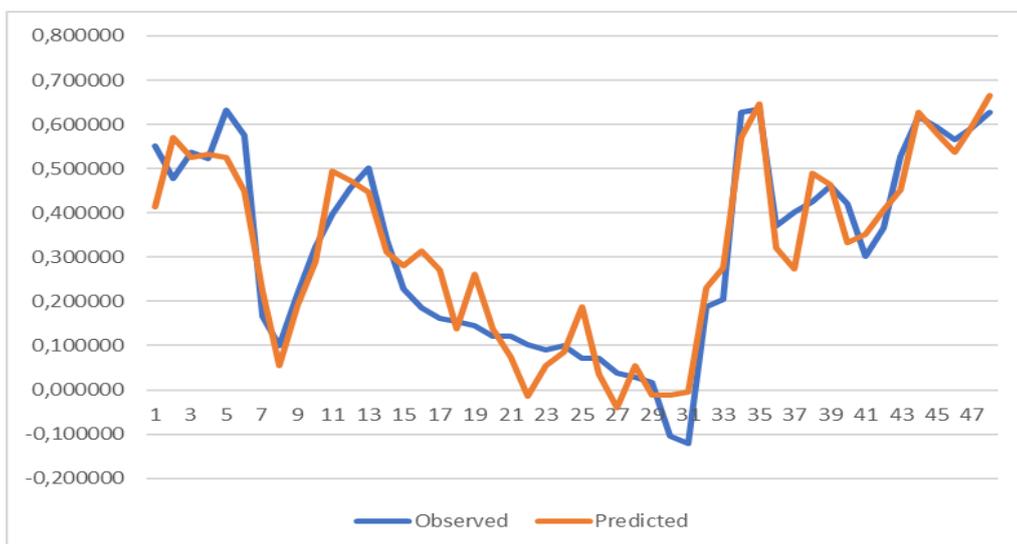


Fig. 27 : Comparaison graphique entre le LOG₁₀(COT) mesuré et estimé par la méthode de RLM utilisant les data brut après la d’élimination des outliers.

la formule de calcul de ce modèle est :

$\text{LOG}_{10}(1/\text{TOC}) = +0.196 \text{ GR} +0.463 \text{ R} +0.699 \text{ S} -0.30 \text{ D} +1.068$

Tab. 16. Illustration de comparaison statistique entre LOG₁₀(COT) mesuré et estimé par la méthode de la RLM utilisant les data brut après la d’élimination des outliers

log10 COT	COT Mesurés	COT Simulés
obs 1	0,552447	0,413937
obs 2	0,476850	0,570745

obs 3	0,536180	0,525083
obs 4	0,522618	0,532265
obs 5	0,632168	0,524321
obs 7	0,576018	0,448417
obs 9	0,165333	0,222355
obs 10	0,098876	0,055711
obs 11	0,215561	0,191268
obs 12	0,323693	0,289092
obs 13	0,396938	0,494493
obs 14	0,455737	0,473976
obs 15	0,500445	0,446262
CC	0,946556	
MSE	0,070227	
MAE	0,057588	

3. LOG₁₀ (DATA)

On a enlevé les observations 06 ;08 ;43et 44

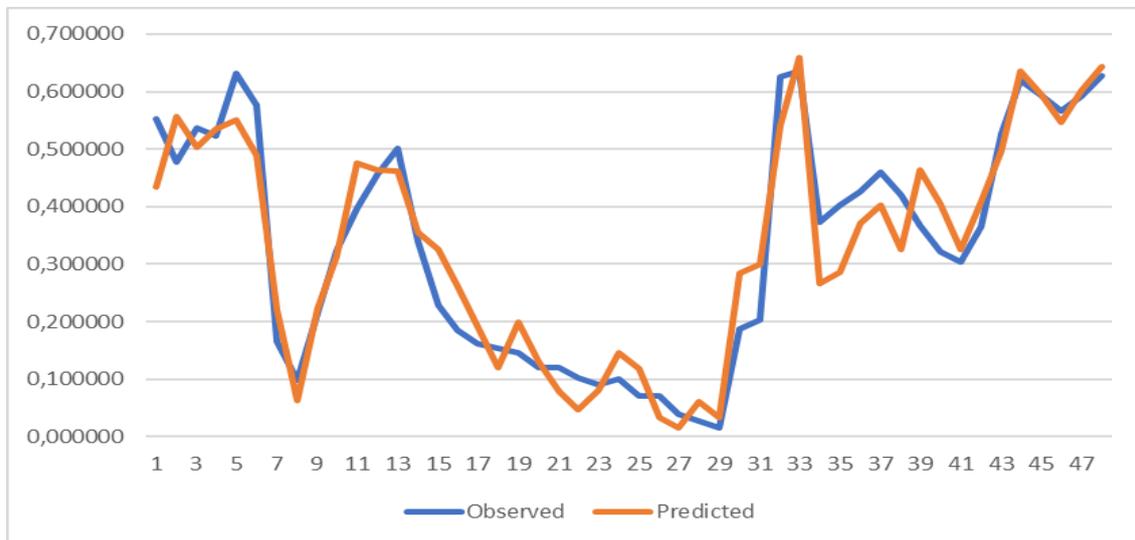


Fig. 28 : Comparaison graphique entre le LOG₁₀(COT)mesuré et estimé par la méthode de RLM utilisant le LOG₁₀(data) après la d’élimination des outliers.

la formule de calcul de ce modèle est :

$$\text{LOG}_{10}(\text{COT}) = 0.241 * \text{LOG}_{10}(\text{GR}) + 0.571 * \text{LOG}_{10}(\text{R}) + 0.687 * \text{LOG}_{10}(\text{S}) - 0.26 * \text{LOG}_{10}(\text{D}) - 2.750$$

Tab. 17. Illustration de comparaison statistique entre $LOG_{10}(COT)$ mesuré et estimé par la méthode de la RLM utilisant le $LOG_{10}(data)$ après la d'élimination des outliers

log10 DATA	COT Mesurés	COT Simulés
obs 1	0,552447	0,433975
obs 2	0,476850	0,556900
obs 4	0,522618	0,535799
obs 5	0,632168	0,550555
obs 7	0,576018	0,489138
obs 9	0,165333	0,219993
obs 10	0,098876	0,062814
obs 11	0,215561	0,222370
obs 12	0,323693	0,313076
obs 13	0,396938	0,475264
obs 14	0,455737	0,464257
CC	0,954518	
MSE	0,059224	
MAE	0,048864	

4. Racine des data

On a enlevé les observations 06 ;08 ;43et 44

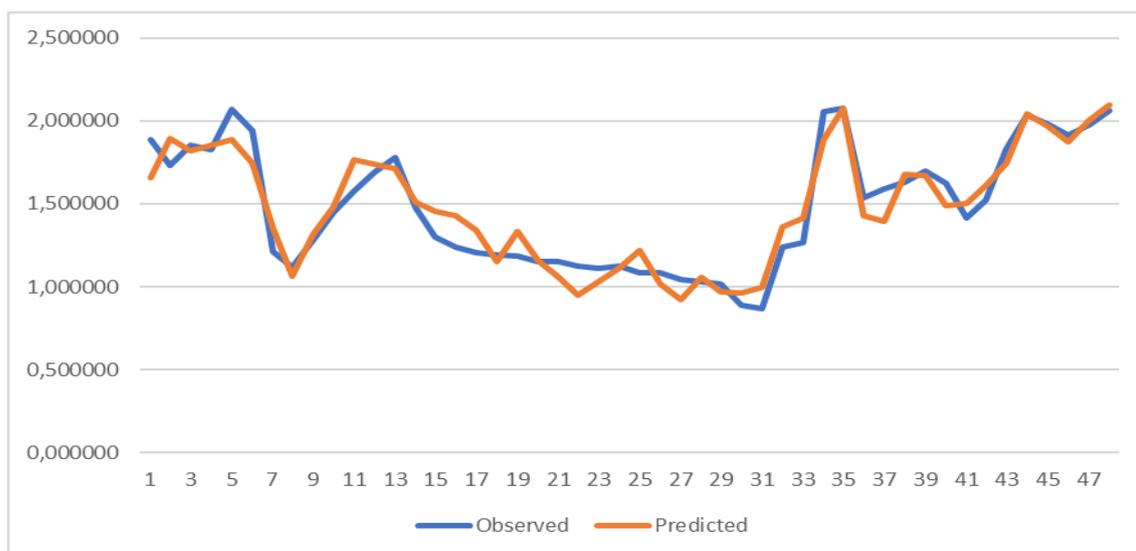


Fig. 29 : Comparaison graphique entre la Racine(COT) mesuré et estimé par la méthode de RLM utilisant la Racine de data après la d'élimination des outliers.

la formule de calcul de ce modèle est :

$$\text{RACINE(COT)} = 0.175 * \text{RACINE(GR)} + 0.511 * \text{RACINE(R)} + 0.595 * \text{RACINE(S)} \\ - 0.389 * \text{RACINE(D)} + 6.802$$

Tab. 18. Illustration de comparaison statistique entre la Racine(COT) mesuré et estimé par la méthode de la RLM utilisant la Racine de data après la d'élimination des outliers

racine data	COT Mesurés	COT Simulés
obs 1	1,888963	1,660085
obs 2	1,731511	1,893693
obs 3	1,853915	1,819281
obs 4	1,825193	1,852466
obs 5	2,070541	1,886568
obs 7	1,940925	1,743484
obs 9	1,209669	1,353272
obs 10	1,120567	1,062675
obs 11	1,281682	1,313797
obs 12	1,451599	1,482240
obs 13	1,579316	1,767907
obs 14	1,689929	1,738356
CC	0,952237	
MSE	0,111441	
MAE	0,091894	

Comparaison

Tab. 19. comparaison statistique entre les différents modèles de la RLM

	1/COT	LOG10 COT	LOG10 DATA	RACINE DATA
CC	0,937386	0,946556	0,954518	0,952237
MSE	0,080024	0,070227	0,059224	0,111441
MAE	0,062791	0,057588	0,048864	0,091894

D'après les résultats et le tableau ci-dessus on remarque que la meilleur méthode est LOG10 DATA avec élimination des outliers de coté d'indices de performances telle que :

L'erreur quadratique moyenne (RMSE) et l'erreur absolue moyenne (MAE) sont très bas (MSE = 0.059, MAE=0.049) et le coefficient de corrélation est proche de 1 (CC = 0.954). Donc les résultats donnés par le modèle LOG10 DATA avec élimination des OUTLIERS sont bons et acceptable.

5 La validation Externe

Cette méthode est de changer dans l'ordre des caractéristiques d'observation d'une manière aléatoire pour essayer de changer entre les valeurs de Trainages et autres de Validation.

Le tableau ci-dessous montre les coefficients de détermination entre différents validations externe (11 validations).

Tab. 20.comparaison statistique entre les coefficients de déterminations de 11 validations externes

	V D	V In	Trainage	Validation	Outliers Supprimé	MSE	R ² Ajusté	Link
LMR 01	TOC	Les diagraphies utilisés	42	0	10	0,1210	0,9015	<u>Linear regression 01</u>
LMR 02	TOC	Les diagraphies utilisés	32	10	10	0,1399	0,8810	<u>Linear regression 02</u>
LMR 03	TOC	Les diagraphies utilisés	32	10	10	0,1502	0,8505	<u>Linear regression 03</u>
LMR 04	TOC	Les diagraphies utilisés	32	10	10	0,1422	0,8821	<u>Linear regression 04</u>
LMR 05	TOC	Les diagraphies utilisés	32	10	10	0,1172	0,9087	<u>Linear regression 05</u>
LMR 06	TOC	Les diagraphies utilisés	32	10	10	0,0941	0,9287	<u>Linear regression 06</u>
LMR 07	TOC	Les diagraphies utilisés	32	10	10	0,1139	0,8989	<u>Linear regression 07</u>
LMR 08	TOC	Les diagraphies utilisés	32	10	10	0,1009	0,9183	<u>Linear regression 08</u>
LMR 09	TOC	Les diagraphies utilisés	32	10	10	0,1140	0,9050	<u>Linear regression 09</u>
LMR 10	TOC	Les diagraphies utilisés	32	10	10	0,1216	0,8862	<u>Linear regression 10</u>
LMR 11	TOC	Les diagraphies utilisés	32	10	10	0,1245	0,8951	<u>Linear regression 11</u>

D'après le tableau ci-dessus en remarque que la meilleure corrélation est celle-ci de LMR 06 de (R^2 Ajusté=0,9287 et MSE=0,0941) voire le tableau de LMR 06 dans L'ANNEX.

Conclusion

Dans ce chapitre on a utilisé différents modèles pour comparer entre le COT mesuré au laboratoire et autre estimé par la méthode de la régression linéaire multiple.

Le meilleur résultat de chaque modèle est

1. Le Transformation du COT : dans ce modèle le $\text{LOG}_{10}(\text{COT})$ est le plus réel.
2. Le Transformation des DATA : dans ce modèle le $\text{LOG}_{10}(\text{DATA})$ est le plus réel.
3. La d'élimination des outliers : dans ce modèle le $\text{LOG}_{10}(\text{DATA})$ après la d'élimination des outliers est le plus réel.
4. La Validation externe : est le LMR 06.

Donc finalement parmi ses modèles On recommande le $\text{LOG}_{10}(\text{DATA})$ avec d'élimination des outliers avec ($\text{CC} = 0.954$; $\text{MSE} = 0.059$ et $\text{MAE}=0.049$) le modèle le plus performant pour estimer le COT.

CONCLUSION GENERALE

et

PERSPECTIVES

CONCLUSION GENERALE

Au cours des dernières années, les chercheurs ont progressé dans l'étude sur l'estimation du COT, alors que les méthodes utilisées dans les recherches se concentrent toujours sur les réponses qualitatives et les formules empiriques. La technique de la régression multiple, avec ça développement rapide et ça exactitude dans la prévision devient un outil très efficace pour prévoir le COT.

Dans notre mémoire on a établi des différents modèles de la régression multiple et Feature Engineering pour la prédiction du COT en utilisant les diagraphies du puits, d'après la visualisation graphique de tous les résultats du COT calculée par les différents modèles avec le COT mesurés, il s'avère très bien que la courbe des $\text{LOG}_{10}(\text{DATA})$ avec D'Elimination des OUTLIERS est presque superposée sur les COT mesurés au laboratoire. À travers ses comparaisons quantitatives et qualitatives des différents modèles d'estimation avec le COT mesuré, il est bien clair que le modèle de $\text{LOG}_{10}(\text{DATA})$ sans OUTLIERS fournit des meilleures estimations du COT confirmant par les indices de performances ($\text{CC}=0,954518$; $\text{MSE}=0,059224$; $\text{MAE}=0,048864$). Grace à cette topologie on peut appliquer ce modèle sur les intervalles non carottés pour récupérer le COT manqué. En plus de cela on peut appliquer le modèle $\text{LOG}_{10}(\text{DATA})$ sans OUTLIERS pour gagner du temps et d'argent par rapport au travaille à l'laboratoire.

Les modèles de la régression linéaire multiple présentent les avantages suivants :

- Le gain économique car ils sont moins couteuses par rapport les analyses de laboratoire.
- Avoir des logs continus de COT avec grande précision.

Ces modèles sont comme toutes méthodes de simulation présentent des inconvénients parmi lesquels :

- La mauvaise interprétation des résultats primaires de COT au laboratoire va provoquer des erreurs au stade des résultats de prédiction.
- une mauvaise calibration des diagraphies ou problèmes des tops géologiques peuvent nous fausser les résultats.

La comparaison entre ses modèles montre que le $\text{LOG}_{10}(\text{DATA})$ sans OUTLIERS s'avèrent plus performante que la transformation de COT et les autres modèles utilisé, et on peut les généraliser sur d'autres caractéristiques géologiques pour un meilleur développement du réservoir.

PERSPECTIVES

Ce travail nous a permis de constater que le modèle de LOG10(DATA) sans OUTLIERS est très utile dans le domaine pétrolier et peut valoriser d'avantage les données disponibles.

alors grâce à LOG10(DATA) sans outliers on peut minimiser et optimiser les dépenses, c'est pas pour chaque puits on doit faire des mesures des diagraphies géochimiques spécifiques qui sont très coûteuses, mais au fur et à mesure on récolte des données des puits voisins, on construit des bases de données assez suffisantes pour élaborer des régressions multiples et performants afin d'estimer le COT avec le moindre d'erreur possible, dans notre cas on a injecté les données diagraphies (Gamma ray, Résistivité, Sonique et Densité) dans l'équation d'un puits voisin et on a validé les résultats trouvés par des mesures de laboratoire.

Références Bibliographiques

Références Bibliographiques

- Abdelmadjid, A. et al. (2019) Evaluation du réservoir schisteux dans le bassin d'Illizi, sud-est de l'Algérie.
- Abdulhamid, A. et al. (2017) 'Determination of the total organic carbon (TOC) based on conventional well logs using artificial neural network', *International Journal of Coal Geology*. Elsevier, 179(May), pp. 72–80. doi: 10.1016/j.coal.2017.05.012.
- Abouelresh, M. O. and Slatt, R. M. (2011) 'Shale Depositional Processes: Example from the Paleozoic Barnett Shale, Fort Worth Basin, Texas, USA', 3(4), pp. 398–409. doi: 10.2478/s13533-011-0037-z.
- Al-Selwi, A. and Joshi, M. (2015) 'Source rock evaluation using total organic carbon (TOC) and the loss-on-ignition (LOI) techniques', *Oil Gas Res*, 1, p. 105.
- Ameur Zaimeche, O., Zeddouri, A., Heddam, S., Kechiched, R., Belksier, M.S., (2014). Modélisation et reconstitution des facies non carottés à l'aide des méthodes statistiques multivariées du réservoir trias argileux greseux inferieur (TAGI) application au champ de sif fatima bassin de – Berkine – Université d'Ouargla, Algérie. 2014
- Amiri Bakhtiar, H., Telmadarreie, A., Shayesteh, M., Heidari Fard, M.H., Talebi, H., Shirband, Z., 2011. Estimating total organic carbon content and source rock evaluation, applying DlogR and neural network methods: Ahwaz and Marun oilfields, SW of Iran. *Pet. Sci. Technol.* 29 (16), 1691-1704.
- Bloechle, J. (2003) 'Réseau de Neurones Artificiels pour la classification des fontes Arabes et la distinction entre la langue Arabe et les langues Latines Table des matières', pp. 1–48.
- Autric, A., Dumesnil, P., 1985. Resistivity radioactivity and sonic transit time logs to evaluate the organic content of low permeability rocks. *Log Anal.* 26 (3), 37-45.
- Decker, A.D., Hill, D.G., Wicks, D.E., 1993. Log-based gas content and resource estimates for the Antrim shale, Michigan Basin. SPE25910. In: *Low Permeability Reservoirs Symposium*. Denver, Colorado, USA. April 26-28.
- Djeriri, Y. (2017) 'Les Réseaux de Neurones Artificiels'.
- Doncescu, A. (no date) 'Les réseaux de neurones artificiels', pp. 1–80.
- Dr Mohamed Said Beghoul (2013) 'Les diagraphies différées'.

Guo, L., Chen, J.F., Miao, Z.Y., 2009. Study and application of a new overlay method of the TOC content. *Nat. Gas. Geosci.* 20 (6), 951-956.

Huang, Z.H., Williamson, M.A., 1996. Artificial neural network modelling as an aid to source rock characterization. *Mar. Pet. Geol.* 13 (2), 277-290.

Heidari, Z., Torres-Verdín, C., Preeg, W.E., 2011. Quantitative method for estimating total organic carbon and porosity, and for diagnosing mineral constituents from well logs in shale-gas formations. In: *SPWLA 52nd Annual Logging Symposium*, Colorado Springs, Colorado, USA, May 14-18.

Huo, Q.L., Zheng, H.S., Fu, L., Ren, Z.G., 2011. The advance of $\Delta\log R$ method and its application in Songliao basin.

Hu, H.T., Lu, S.F., Liu, C., Wang, W.M., Wang, M., Li, J.J., Shang, J.H., 2011. Models for calculating organic carbon content from logging information: comparison and analysis. *Acta Sedimentol. Sin.* 29 (6), 1199-1205.

Ibrahim, B. and Abdelaziz, N. (2018) 'Apport des données de diagraphies dans la caractérisation des réservoirs shale gaz : Cas du Lower Barnett USA'.

Ifpen (2012) 'Les hydrocarbures non conventionnels : évolution ou révolution ?'

Ingleby, H. R. and Crowe, T. G. (2001) 'Neural network models for predicting organic matter content in Saskatchewan soils', *Outlook*, 1(2), p. 3.

Ireland, P. E. (2012) 'Le Barnett Shale est entré dans l ' histoire', pp. 10–12.

(Philippe MICHEL)file:///C:/Users/Migliore/Desktop/master%202/regrission%20multiple/donnees%20ouiafi/Transformation%20angulaire%20-%20[Trucs%20pour%20LaTeX%20&%20R].html

Journal of Jilin University (Earth Science Edition). 41(2), 586–591.

Kamali, M.R., Mirshady, A.A., 2004. Total organic carbon content determined from well logs using $\Delta\log R$ and Neuro 583 Fuzzy techniques. *J.f Pet. Sci. Eng.* 45, 141-148

Li, H., Wang, H., 2016. Investigation of eccentricity effects and depth of investigation of azimuthal resistivity LWD tools using 3D finite difference method: *Journal of Petroleum Science and Engineering*. 143(7), 211-225.

Meyer, B.L., Nederlof, M.H., 1984. Identification of source rocks on wireline logs

by density/resistivity and sonic transit/resistivity crossplots. *AAPG Bull.* 68, 121-129.

Schmoker, J., Hester, T., 1983. Organic carbon in bakken formation, United States portion of williston basin. *Am. Assoc.Pet. Geol. Bull.* 67, 2165-2174.

O.SERRA (1985) 'Diagraphies Différées(bases de l'interprétation), Acquisition des données diagraphiques', pp. 14–34.

Ouadfeul, S. and Aliouane, L. (2015) 'Total Organic Carbon Prediction from well Logs Data using the Multilayer Perceptron with Levenberg Marquardt Algorithm with an Application to Shale Gas Reservoirs'.

Rouaud, R. V. – G. M. – T. (2013) 'Hydrocarbures de roche-mère'.

Schmoker, J. W. and Hester, T. C. (1983) 'Organic carbon in Bakken formation, United States portion of Williston basin', AAPG bulletin. American Association of Petroleum Geologists, 67(12), pp. 2165–2174.

Schmoker, J. W., 1979, Determination of organic content of Appalachian Devonian shale from formation-density logs: AAPG Bulletin, 63(9), 1504-1509.

Shi, X., Wang, J., Liu, G., Yang, L., Ge, X.M., Jiang, S., 2016. Application of extreme learning machine and neural networks in total organic carbon content prediction in organic shale with wire line logs. J.Nat. Gas Sci. Eng. 33,687-702.

Théorêt, É. (2015) 'EXPLOITATION NON CONVENTIONNELLE : ANALYSE CRITIQUE DU RÈGLEMENT SUR LE PRÉLÈVEMENT DES EAUX ET LEUR PROTECTION', Biomass Chem Eng, 49(23–6).

TRANIER, J. (2006) 'Modélisation stratigraphique de la distribution et de la préservation de la matière organique en milieu marin profond. Cas d'une marge à sédimentation pélagique : système d'upwelling côtier du Benguela (Namibie, Afrique du Sud Ouest).'

US Geological Survey (no date) 'Les hydrocarbures conventionnels et non conventionnels'.

WikiStat (no date) 'Introduction à la régression multiple', pp. 1–6.

Zhao, Z. et al. (2010) 'Using artificial neural network models to produce soil organic carbon content distribution maps across landscapes', Canadian journal of soil science. NRC Research Press, 90(1), pp. 75–87.

ANNEXE

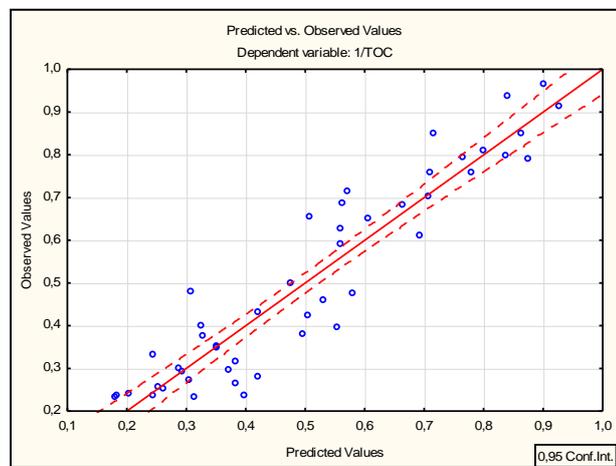
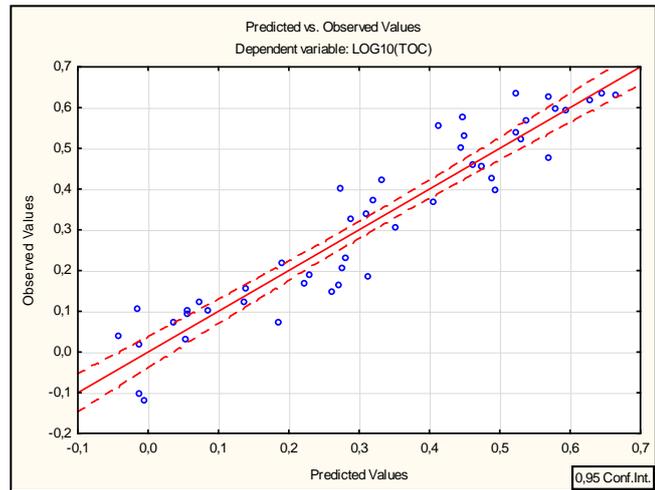
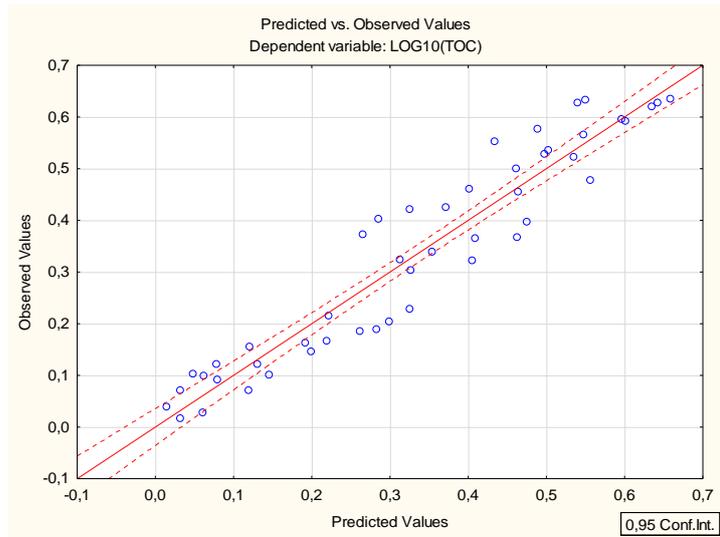
ANNEXE

Observation	TOC	Pred(TOC)	Residual	
Obs1	0,7570	0,7479	0,0091	Traning
Obs2	0,7850	0,8778	-0,0928	Traning
Obs4	1,0654	1,2932	-0,2278	Traning
Obs5	1,0935	0,7816	0,3119	Traning
Obs7	1,1776	1,2459	-0,0683	Traning
Obs8	1,2336	1,1009	0,1327	Traning
Obs9	1,2557	1,2866	-0,0310	Traning
Obs10	1,2617	1,1159	0,1458	Traning
Obs11	1,3204	1,2389	0,0815	Traning
Obs12	1,3209	1,5952	-0,2743	Traning
Obs13	1,4001	1,7088	-0,3088	Traning
Obs14	1,4270	1,4173	0,0096	Traning
Obs17	1,5421	1,4547	0,0873	Traning
Obs19	1,6427	2,0493	-0,4066	Traning
Obs20	1,6901	2,1845	-0,4944	Traning
Obs24	2,3208	2,4144	-0,0937	Traning
Obs25	2,3551	1,8155	0,5396	Traning
Obs26	2,4942	3,0323	-0,5380	Traning
Obs27	2,6355	2,2580	0,3775	Traning
Obs28	2,6636	3,0800	-0,4164	Traning
Obs29	2,8559	3,1825	-0,3267	Traning
Obs30	2,8879	2,4891	0,3987	Traning
Obs31	2,9981	3,3039	-0,3058	Traning
Obs33	3,3313	3,3040	0,0273	Traning
Obs34	3,4370	3,2462	0,1908	Traning
Obs35	3,5682	3,0841	0,4840	Traning
Obs36	3,6793	3,3328	0,3465	Traning
Obs37	3,9057	3,8842	0,0214	Traning
Obs38	3,9340	3,7380	0,1959	Traning
Obs39	4,1604	3,8038	0,3566	Traning
Obs40	4,2453	4,2510	-0,0057	Traning
Obs42	4,3178	4,4440	-0,1262	Traning
Obs3	1,0374	0,9087	0,1287	validation
Obs6	1,1776	1,8167	-0,6391	validation
Obs15	1,4537	2,0111	-0,5575	validation
Obs16	1,5334	1,5093	0,0241	validation
Obs18	1,5981	1,9710	-0,3729	validation
Obs21	2,0943	2,8167	-0,7224	validation
Obs22	2,1071	1,6346	0,4725	validation
Obs23	2,1818	2,5754	-0,3937	validation
Obs32	3,1655	3,0234	0,1421	validation
Obs41	4,2871	3,8028	0,4844	validation

Tableau de Linear regression 06
Le meilleur modèle de la validation externe

Observation	TOC	Pred(TOC)	Residual	
Obs3	1,0374	0,9444	0,0930	Traning
Obs4	1,0654	1,3697	-0,3043	Traning
Obs5	1,0935	0,7838	0,3096	Traning
Obs7	1,1776	1,3193	-0,1418	Traning
Obs8	1,2336	1,0161	0,2175	Traning
Obs10	1,2617	1,0537	0,2080	Traning
Obs11	1,3204	1,1566	0,1638	Traning
Obs12	1,3209	1,5054	-0,1845	Traning
Obs13	1,4001	1,6929	-0,2928	Traning
Obs14	1,4270	1,3542	0,0728	Traning
Obs15	1,4537	1,9489	-0,4952	Traning
Obs16	1,5334	1,5858	-0,0524	Traning
Obs17	1,5421	1,4625	0,0796	Traning
Obs18	1,5981	1,9236	-0,3254	Traning
Obs19	1,6427	2,1035	-0,4608	Traning
Obs21	2,0943	2,7996	-0,7053	Traning
Obs22	2,1071	1,8986	0,2085	Traning
Obs23	2,1818	2,4999	-0,3182	Traning
Obs24	2,3208	2,3574	-0,0367	Traning
Obs25	2,3551	1,8035	0,5516	Traning
Obs26	2,4942	3,0215	-0,5273	Traning
Obs27	2,6355	2,1984	0,4371	Traning
Obs30	2,8879	2,4991	0,3887	Traning
Obs31	2,9981	3,2604	-0,2623	Traning
Obs32	3,1655	2,9605	0,2051	Traning
Obs33	3,3313	3,2703	0,0610	Traning
Obs34	3,4370	3,2247	0,2123	Traning
Obs37	3,9057	3,9511	-0,0454	Traning
Obs38	3,9340	3,7522	0,1818	Traning
Obs39	4,1604	3,8245	0,3359	Traning
Obs40	4,2453	4,3414	-0,0961	Traning
Obs41	4,2871	3,7648	0,5223	Traning
Obs1	0,7570	0,7387	0,0183	validation
Obs2	0,7850	0,8024	-0,0173	validation
Obs6	1,1776	2,0250	-0,8475	validation
Obs9	1,2557	1,3633	-0,1076	validation
Obs20	1,6901	2,1465	-0,4564	validation
Obs28	2,6636	2,9986	-0,3351	validation
Obs29	2,8559	3,1110	-0,2552	validation
Obs35	3,5682	2,9811	0,5871	validation
Obs36	3,6793	3,3146	0,3647	validation
Obs42	4,3178	4,3546	-0,0369	validation

Tableau de Tableau de Linear regression 07
D'un modèle de la validation externe



Les valeurs observés en fonction des valeurs estimés
Cas de d'élimination des outliers

الملخص

يقدر إجمالي الكربون العضوي (COT) ثراء صخر المصدر بالمادة العضوية ومدى استعداده لتوليد الهيدروكربونات. يؤهل محتوى الكربون العضوي الكلي إمكانات صخرة المصدر. إن أفضل مصدر لتقدير الكربون العضوي الكلي هو القياس المباشر من العينات الصخرية في المختبر، ولكن بسبب المشاكل الجيولوجية أو التقنية أثناء عملية الحفر، فإن استرداد العينات الصخرية ليس دائمًا كاملاً، بالإضافة إلى ذلك، عملية تحديد الكربون العضوي الكلي مكلفة جداً وتستغرق وقتاً طويلاً. في عملنا على حقل الغاز الصخري في بارنيت، استخدم تطبيق الانحدار المتعدد وميزة الهندسة وأشكالهما المتعددة على سجلات الآبار الكلاسيكية (أشعة جاما، المقاومة، سونيك والكثافة) من أجل تحسين النتائج، وقد أعطت هذه الأشكال دقة كبيرة $(CC=0.88 ; 0.95, RMSE=0.54;0.06, MAE=0.40 ; 0.05)$ ، بفضل طريقة اللوغاريتم العشري للبيانات مع حذف القيم الشاذة التي تتمتع بأداء رائع، كما يمكننا حل مشكلة 51 نقطة غير مقاسة، وتقليل التكلفة وتوفير الوقت.

كلمات المفتاح:

الكربون العضوي الكلي - مشاكل الاسترجاع - ميزة الهندسة- التحسين- الانحدار المتعدد - تسجيلات الآبار - بارنيت

RESUMÉ

Le carbone organique total COT évalue la richesse de la roche mère en matière organique et sa prédisposition à générer des hydrocarbures. La teneur en COT permet de qualifier le potentiel de la roche mère. La meilleure source d'estimation du COT est la mesure directe à partir des carottes dans le laboratoire, mais à cause aux problèmes géologiques ou techniques pendant l'opération de carottage la récupération des carottes n'est pas toujours intégrales, En outre, le dosage du COT est coûteux et prend du temps.

Dans notre travail sur champ de gaz de schiste de Barnett, on a utilisé la régression multiple et Feature Engineering en différents modèles pour but d'amélioré les résultats, ses modèles ont été appliqués sur des diagraphies classiques (Gamma ray, Résistivité, Sonique et Densité), elles donnent des grandes précisions $(CC=0.88 ; 0.95, RMSE=0.54 ; 0.06, MAE=0.40 ; 0.05)$. Grâce au modèle de LOG10(DATA) sans OUTLIERS qui a une grande performance, on peut résoudre le problème de 51 points non mesurés, minimiser le cout et le gain du temps.

Mots-Clés :

COT, problèmes de récupération, Feature engineering, amélioration, régression linière multiple, diagraphies, Barnett.

ABSTRACT

Total organic carbon TOC values the richness of the source rock in organic matter and its predisposition to generate hydrocarbons. The TOC content qualifies the potential of the source rock. The best source of TOC estimation is the direct measurement from the cores in the laboratory, but because of the geological or technical problems during the coring operation the recovery of the cores is not always integral, In addition, the dosage TOC is expensive and time consuming. In our work on Barnett's shale gas field, multiple regression and Feature Engineering and it's models were applied on classical logs (Gamma ray, Resistivity, Sonic and Density) to improve the results, they gives great precision $(CC=0.88 ; 0.95, RMSE=0.54;0.06, MAE=0.40 ; 0.05)$, Thanks to the LOG10(DATA) without OUTLIERS which has a great performance, we can solve the problem of 51 unmeasured points, minimize cost and time saving.

Keywords :

TOC, recovery problems, multiple regression, Feature Engineering, Improvement, logging, Barnett