HAL
archives-ouvertes.fr

Modélisation de la viscosité des fluides de gisement : Apport de la PLS bootstrap et des réseaux de neurones

P. H. Gayon, A. Pina, I. Ahmed, A. Faraj

▶ **To cite this version:**

P. H. Gayon, A. Pina, I. Ahmed, A. Faraj. Modélisation de la viscosité des fluides de gisement : Apport de la PLS bootstrap et des réseaux de neurones. Oil & Gas Science and Technology - Revue d'IFP Energies nouvelles, Institut Français du Pétrole, 2008, 63 (5), pp.629-643. 10.2516/ogst:2008024 . hal-02002039

HAL Id: hal-02002039
https://hal-ifp.archives-ouvertes.fr/hal-02002039

Submitted on 31 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Oil & Gas Science and Technology – Rev. IFP, Vol. 63 (2008), No. 5, pp. 629-643
Copyright © 2008, Institut français du pétrole
DOI: 10.2516/ogst:2008024

Modélisation de la viscosité des fluides de gisement : Apport de la PLS bootstrap et des réseaux de neurones

P.H. Gayon[2], A. Pina[1*], I. Ahmed[3] et A. Faraj[1]

1 Institut français du pétrole, IFP, 1-4 avenue Bois-Préau, 92852 Rueil-Malmaison Cedex - France
2 Perenco, 23 rue Dumont d'Urville, 75116 Paris - France
3 Inserm, U780, 16 avenue Paul Vaillant-Couturier, 94807 Villejuif - France
e-mail: rprodeng@cd.perenco.com - annabelle.pina@ifp.fr - ismail.ahmed@inserm.fr - abdelaziz.faraj@ifp.fr

* Corresponding author

Résumé — La régression linéaire Partial Least Squares (PLS) avec ré-échantillonnage par bootstrap et la régression non-linéaire par réseaux de neurones ont été utilisées pour construire des modèles de prédiction de la viscosité de fluides de gisement en fonction de leur composition, de la température et de la pression. Les modèles statistiques résultants permettent de prédire des valeurs de viscosité s'étendant de 0,21 cP à 10000 cP.

Abstract* — *Viscosity Modelling for Reservoir Fluids: Contribution of Partial Least Square with Bootstrap Techniques and Neural Networks — *Partial Least Squares linear regression (PLS) with bootstrap sampling, and Multi Layer Perceptron (MLP) non-linear regression were used to elaborate new viscosity models for estimating crude oil viscosity using fluid composition, temperature, and pressure. Statistical models obtained concern values from 0.21 cP to 10000 cP.*

INTRODUCTION

La viscosité des fluides de gisement est une propriété physique très importante qui contrôle et influence les migrations et les flux d'huile au sein d'un bassin ou des lignes de production. Ce paramètre représente un des facteurs essentiels des simulateurs de bassin, de réservoir et de production. En effet, lors de la récupération et du traitement du brut, il est nécessaire de connaître la viscosité dans une large gamme de pression et de température, mais les mesures réalisées sur les fluides de gisement sont délicates à réaliser, il n'est donc pas possible de les effectuer sur l'ensemble du domaine de température et de pression. C'est pourquoi le développement de modèles de viscosité semble pertinent pour réduire les coûts des études expérimentales, mais aussi les coûts d'exploitation. En effet, lors de projets de développement de champs pétrolifères, la connaissance de la viscosité permet, d'une part, une meilleure estimation des réserves en place et, d'autre part, un meilleur dimensionnement des installations de surface et de fond.

Contrairement aux gaz, la théorie de la viscosité des liquides a été peu développée. En effet, les forces intermoléculaires entre chaque molécule, leur structure et le degré de désordre de l'ensemble des molécules sont tels qu'il n'existe pas de méthode théorique pour prévoir exactement la viscosité des liquides. Nous pouvons cependant rencontrer plusieurs types de méthodes de prédiction dans la littérature :

- méthodes empiriques,
- modèles spécifiques,
- modèles des états correspondants,
- modèles basés sur les équations d'état,
- modèles compositionnels.

Dans ce travail, une approche statistique de prédiction de la viscosité est envisagée ; l'Analyse Factorielle Discriminante (AFD), la régression Partial Least Squares (PLS) et les réseaux de neurones sont utilisés pour construire un modèle statistique de viscosité à partir des variables température, pression et composition.

En première partie et à titre de comparaison avec les méthodes de calcul proposées dans cet article, nous présenterons les différents modèles de viscosité rencontrés classiquement dans la littérature. La deuxième partie sera consacrée à la description de l'acquisition des données ainsi qu'à l'étude de leur domaine de validité. Ces données seront utilisées dans la troisième partie pour la détermination de modèles statistiques. Pour finir, une comparaison du modèle statistique avec les modèles rencontrés dans la littérature basée sur les données non utilisées pour construire le modèle permettra d'évaluer cette nouvelle approche. Dans l'annexe figure une description succincte des méthodes statistiques utilisées : régression Partial Least Squares, analyse factorielle discriminante et réseaux de neurones ainsi que la présentation des paramètres déterminés.

1 MÉTHODES DE PRÉDICTION DE LA VISCOSITÉ

La viscosité des fluides de gisement possède une échelle de variation importante. Température, pression, composition du fluide sont des facteurs déterminants qui font varier de manière importante la viscosité du brut. Ces grandeurs changent au fur et à mesure de la déplétion du champ ; obtenir une représentation de la viscosité du brut pendant la vie du réservoir est alors difficile. C'est pourquoi, plusieurs types de modèles de prédiction ont été envisagés pour déterminer la viscosité de ces fluides de gisement.

1.1 Corrélations empiriques

Parmi les lois empiriques, les plus utilisées permettent d'estimer la viscosité des huiles de gisement à partir d'un nombre limité de facteurs mesurables comme la température, la pression, la pression de bulle, la densité de l'huile et du gaz et le rapport entre le volume de gaz dissous dans l'huile dans les conditions standards et le volume de brut de stockage dans les mêmes conditions (Rs). Les corrélations empiriques sont de quatre types ; elles permettent de modéliser la viscosité :

- à la pression atmosphérique, où le fluide est appelé huile morte ou huile de stockage et ne contient plus de gaz ;
- à la pression de bulle, où l'huile est dite saturée ;
- au-dessus de la pression de bulle, on parle alors d'huile sous-saturée de composition constante ;
- en dessous de la pression de bulle, l'huile est dite sur-saturée et de composition variable causée par la chute de pression dans le réservoir et la libération des gaz dissous. Le fluide devient alors de plus en plus lourd provoquant un accroissement de sa viscosité.

À titre d'exemple, on peut citer les modèles empiriques de Beggs et Robinson (1975) [1], Petrosky et Farshad (1995) [2], Labedi (1992) [3] et Khan *et al*. (1987) [4].

1.2 Modèles spécifiques

Certaines méthodes permettent de calculer les effets de la température et/ou de la pression sur la viscosité d'un fluide. Ces méthodes nécessitent une viscosité de brut de stockage en entrée du modèle, c'est pourquoi nous les distinguons sous le terme de modèles spécifiques. Les prédictions de l'effet de la température sont basées sur l'équation d'Andrade (1934) [5], tandis que les prédictions de l'effet de la pression reposent sur l'équation de Tait (1888) [6]. L'intérêt de ces deux types de prédiction est réduit car elles nécessitent toutes deux un ajustement préalable des paramètres ou un nombre important de données expérimentales. En effet, pour la température, comme pour la pression, ces prédictions sont spécifiques aux fluides pour lesquels elles ont été déterminées. On peut citer les travaux de Dymond et Oye (1994) [7] pour la température et Tanaka *et al*. (1991) [8], ainsi que Assael *et al*. (1991) [9] pour la pression.

Des prédictions simultanées de l'effet de la température et de la pression ont été envisagées par Comunas *et al*. (2001) [10] et Kanti *et al*. (1989) [11], permettant ainsi de combiner les lois d'Andrade et de Tait. En effet, ces modèles utilisent en entrée de la loi en température la viscosité du brut de stockage, et une équation de type Tait pour calculer l'effet de la pression sur la viscosité de fractions liquides.

1.3 Modèles des états correspondants

Les modèles à un état de référence sont basés sur le théorème des états correspondants proposé par Van der Waals dans sa thèse (1873) [12] pour la détermination du facteur de compressibilité. Son principe est le suivant : une propriété d'un fluide peut être prédite à partir de celle d'un autre fluide pris comme référence en passant par l'intermédiaire d'une pression réduite (fraction de la pression nominale sur la pression critique) et d'une température réduite. Cette méthode appliquée à la viscosité permet de prédire la viscosité d'un brut à partir de la température, de la pression, des propriétés pseudocritiques du mélange et de la viscosité d'un fluide de référence évaluée à une pression et une température de référence. Dans le monde pétrolier, cette méthode a été envisagée par Pedersen et Fredenslund (1984) [13] qui fait intervenir en plus un paramètre α pour prendre en compte les masses volumique et molaire du fluide étudié. Un second modèle faisant intervenir le principe des états correspondants est le modèle de Lohrenz *et al*. (1964) [14] proposé dans la majeure partie des simulateurs de réservoir.

1.4 Modèles basés sur les équations d'état

Basés sur l'analogie entre les relations pression-volume-température *P-V-T* et pression-viscosité-température *P*-η-*T*, des modèles ont été développés à partir des équations d'état pour déterminer la viscosité de mélange de gaz et d'huile. On peut citer comme modèles de type équation d'état celui de Little et Kennedy (1968) [15] qui reprend l'équation de Van der Waals et, plus récemment, celui de Guo *et al*. (1997) [16], développé à partir des équations d'état de Patel-Teja (1982) [17] et Peng-Robinson [18].

Une autre approche des équations d'état a été envisagée par Quinones-Cisneros *et al*. (2001, 2004) [19, 20] à travers la *friction theory*. En effet, l'expression de la viscosité d'un fluide de gisement peut se scinder en deux termes, un premier qui concerne les gaz dissous et un deuxième terme de viscosité de friction. Ce dernier est lié aux coefficients de pression d'attraction et de répulsion de type équation d'état de Van der Waals, tels que les équations d'état de Peng Robinson (PR) et Soave Redlich Kwong (SRK) [18]. Combinées à ces équations, des lois de mélange basées sur la composition du fluide et un paramètre d'ajustement permettent la résolution de cette méthode.

1.5 Modèles de volume libre

L'approche du volume libre suppose qu'un liquide est constitué de sphères dures dans lequel il y a possibilité de diffusion grâce au volume libre entre les sphères. Cohen et Turnbull (1959) [21] et Doolittle (1951) [22] ont appliqué cette approche à la modélisation de la viscosité et plus récemment, Allal *et al*. (2001) [23] et Boned *et al*. (2004) [24].

1.6 Modèles compositionnels

Habituellement, la viscosité du fluide de gisement est mesurée de manière isotherme à différentes pressions. Cependant, pendant la déplétion du champ, la composition du fluide peut changer si la pression de saturation est atteinte dans le réservoir, ainsi un modèle compositionnel peut être utilisé pour actualiser la valeur de la viscosité non seulement en fonction de la pression et de la température, mais aussi en fonction de la composition du mélange.

Un modèle récent établi par Elsharkawy et al. (2003) [25] permet, à partir de la composition du mélange (jusqu'au C7+), de la densité, de la masse molaire, de la température et de la pression, de calculer la viscosité du fluide dans différentes conditions. Un autre modèle, W3BH [26, 27] développé dans la thèse de Werner (1996) [28], combine la loi de mélange de Grunberg et Nissan (1949) [29] et le modèle de Kanti *et al*. (1989) [11]. Comme Elsharkawy, ce modèle W3BH permet de prédire la viscosité d'un fluide en fonction de sa composition, sa pression et sa température. Les facteurs d'entrée de ce modèle décrits dans le paragraphe suivant ont servi de base de développement pour le modèle statistique étudié.

2 DONNÉES EXPÉRIMENTALES

2.1 Paramètres de modélisation

L'étude expérimentale effectuée dans la thèse de Werner (1996) [28] sur la viscosité des fluides pétroliers riches en produits lourds permet de regrouper les constituants d'un fluide en pseudo-composants suivant leur influence sur sa viscosité. Ainsi, quatre fractions ressortent de cette étude : les différents composants gazeux peuvent être rassemblés en une seule fraction « Gaz » (non hydrocarbure tel que le CO_2 et hydrocarbures C_1-C_5) ; les hydrocarbures linéaires et cycliques peuvent également être regroupés en un seul pseudo-composant « C_6-C_{20} » ; alors que les constituants lourds C_{20+} sont divisés en deux fractions : une première « C_{20+} saturés + aromatiques + résines » et une seconde composée uniquement de « C_{20+} asphaltènes » rendant l'influence déterminante des asphaltènes sur la viscosité dynamique des fluides de gisement. Les facteurs déterminants du modèle étudié sont alors au nombre de six :

- température, *T*,
- pression, *P*,

– fraction massique Gaz W_{gaz} (hydrocarbures C_1-C_5 et non hydrocarbures),
– fraction massique $W_{C6\text{-}C20}$,
– fraction massique $W_{C20+\ S\text{-}A\text{-}R}$ (Saturés, Aromatiques, Résines),
– fraction massique $W_{C20+\ asph}$ (asphaltènes).

2.2 Acquisition de données

Les techniques expérimentales utilisées au cours de ce travail pour l'acquisition des données se subdivisent en deux parties. La première étape a consisté à analyser les fluides et la seconde à mesurer leur viscosité.

2.2.1 Analyse des fluides de gisement

Cette étape est fondamentale pour établir un modèle compositionnel. En effet, la composition des fluides qui va permettre de construire le modèle doit être la plus précise possible. Tout d'abord, une analyse des huiles par chromatographie en phase gazeuse (CPG) permet d'identifier la composition des bruts de stockage jusqu'au C_{20}, ensuite une distillation pour la séparation des fractions C_{20+}/C_{20-} est effectuée pour permettre un fractionnement des Saturés, Aromatiques, Résines et Asphaltènes (SARA) de la coupe C_{20+} par chromatographie en phase liquide HPLC. La quantité d'asphaltènes est alors mesurée, selon un protocole opératoire optimisé de la norme NFT 60-115, et séparée des autres constituants SARA de la fraction C_{20+}. À l'issue de ces analyses, la composition de chaque fluide étudié est disponible pour la modélisation.

2.2.2 Mesure de la viscosité

Basé sur le principe d'une bille roulante dans un tube calibré incliné, le viscosimètre à bille roulante est un équipement qui permet de mesurer des viscosités en pression et en température (Hubbard et Brown, 1943 [30] ; Stanislawski et Luft, 1987 [31] ; Sawamura *et al*., 1990 [32] ; Barrufet *et al*., 1999 [33] ; Pensado *et al*., 2005 [34]). Notre équipement couvre des températures allant de 25 °C à 150 °C et des pressions allant de 1 bar à 700 bar. Il contient un chronomètre mesurant le temps de roulement de la bille entre deux détecteurs magnétiques proportionnel à la viscosité du fluide. La connaissance d'une constante d'étalonnage C et des masses volumiques de la bille utilisée et du fluide étudié permet de calculer la viscosité du fluide en pression et en température à l'aide de l'équation suivante :

$$\eta_{P,T} = C.(\rho_{bille} - \rho_{fluide}).\Delta t$$

Le fonctionnement théorique de cet appareil a été décrit de manière détaillée par Hubbard et Brown (1943) [30]. En écrivant de manière rigoureuse le bilan des forces, ils démontrent que la constante C peut être théoriquement calculée à partir des dimensions physiques de l'équipement et de son angle d'inclinaison, lorsque l'acquisition du temps de roulement est effectuée en régime laminaire. Ils obtiennent l'expression suivante pour la constante C :

$$C = \frac{5\pi}{42} Kg \sin\theta \frac{d(D+d)}{L}$$

avec :
d le diamètre de la bille
D le diamètre du tube
K une constante uniquement fonction du rapport d/D
θ l'angle que fait le tube par rapport à l'horizontal
L la longueur parcourue par la bille.

En raison des caractéristiques propres à chaque appareil, il est cependant nécessaire de déterminer la constante C par calibration à l'aide d'étalons de viscosité. En réalisant cet étalonnage sur deux billes de dimensions différentes, nous avons observé un écart relatif par rapport à la constante théorique inférieure à 10 %. Enfin, Hubbard et Brown (1943) [30] ont montré que la constante C devait être déterminée en fonction de la température en raison de la dilatation du tube et de la bille tandis que la pression n'avait pas d'effet sur cette constante. Dans un premier temps, dans le cas de notre équipement, les déterminations de la constante d'étalonnage à différentes températures à pression atmosphérique ont montré que la constante C n'était pas influencée par la dilatation du tube et de la bille. Ce résultat peut être imputé au fait que le tube et les billes utilisées sont réalisés en acier inoxydable, matériau dont le coefficient de dilatation linéaire est très faible (compris entre 10^{-6} et 16.10^{-6} K^{-1} entre 20 et 200 °C selon le type d'acier). Dans un second temps, des mesures réalisées en pression avec des fluides de viscosités connues ont montré que la constante d'étalonnage n'était pas influencée par la pression comme cela est le cas pour les équipements utilisés par Hubbard et Brown (1943) [30] et Sawamura *et al*. (1990) [32] et a contrario de Barrufet *et al*. (1999) [33] et Pensado *et al*. (2005) [34].

Après étalonnage de l'équipement, les mesures ont été réalisées en mesurant six temps d'écoulement pour chaque condition (P, T). Ces mesures ont été moyennées pour le calcul de la viscosité.

2.2.3 Mesure de la masse volumique du fluide

Le protocole utilisé pour déterminer la masse volumique a déjà été décrit dans la littérature (Arnaud, 1995 [35]). Brièvement, la masse volumique du fluide (ρ_{fluide}) est mesurée préliminairement au cours d'une étude PVT à l'aide d'une cellule permettant de prélever une partie du fluide en pression et température. La mesure de la masse volumique est alors obtenue en connaissant le volume introduit dans la cellule à l'aide de la cellule volumétrique et la masse par pesée avant et après introduction.

2.2.4 Incertitude des mesures

L'incertitude sur la masse du fluide obtenue par pesée est égale à 10^{-3} g. L'incertitude sur le volume du fluide est estimée à 0,1 cm^3. L'incertitude maximum sur la masse volumique du fluide est estimée égale à 5.10^{-3} g/cm^3. L'incertitude sur la masse volumique de la bille est calculée à l'aide de l'incertitude sur la masse de la bille et son diamètre. La pression a été mesurée à l'aide d'un capteur HBM P3MB /1000 bars. Son incertitude a été déterminée à l'aide d'une jauge à poids mort de type DH Budenberg égale à +/– 1bar. La température a été mesurée à l'aide d'une sonde Pt100 de classe A. Son incertitude, mesurée à l'aide d'une sonde de température de référence, est égale à 0,35 °C. L'incertitude sur la viscosité est calculée pour chaque mesure en fonction de l'incertitude sur les masses volumiques, l'incertitude du temps d'écoulement et l'incertitude de la constante *C* déterminée au cours de l'étalonnage. Sur l'ensemble des mesures acquises, l'incertitude sur la viscosité est inférieure à 10 %.

2.3 Données expérimentales disponibles

Une base de données interne et confidentielle de 196 points expérimentaux de température, pression, composition et viscosité déterminées, correspondant à 16 fluides d'origines différentes (Émirats Arabes Unis, Arabie Saoudite, Orénoque, Canada, Mexique), a été utilisée dans cette étude pour développer un nouveau modèle de viscosité. La figure 1 et le tableau 1 présentent le diagramme d'occurrence de cette base de données et son domaine de validité.

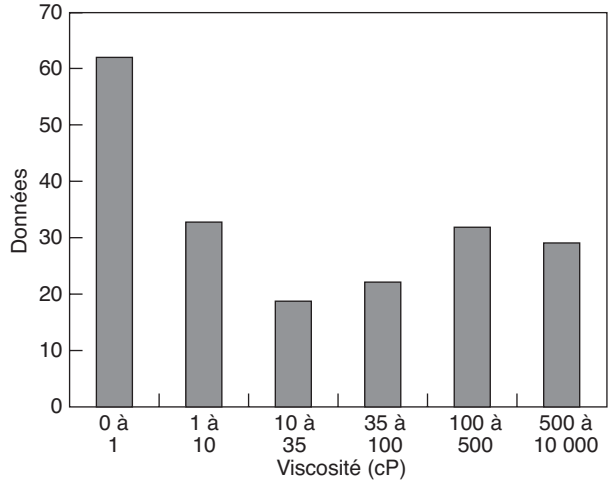

Figure 1

Histogramme de la viscosité des 196 données expérimentales.
Frequency distribution of viscosity variables for 196 experimental data points.

TABLEAU 1

Domaine de validité des données
Ranges of input variable

196 données	W_{gaz} (%)	$W_{C6\text{-}C20}$ (%)	$W_{C20+\ SAR}$ (%)	$W_{C20+\ asph}$ (%)	T (°C)	P (bar)	η_{exp} (cp)
min	0	12,53	27,63	0,06	26,3	14	0,21
max	22,16	51,03	78,13	13	150	626	8743

3 MODÉLISATION DE LA VISCOSITÉ À L'AIDE DE MÉTHODES STATISTIQUES

3.1 Introduction

Cette section est consacrée à la description de la démarche et des résultats obtenus pour la modélisation statistique de la viscosité de fluides de gisement. L'originalité de la méthodologie utilisée ici réside dans l'association de méthodes classiques de régression (Partial Least Squares, réseaux de neurones) avec des méthodes d'échantillonnage (par bootstrap, par validation croisée). L'échantillonnage des données a deux avantages. Tout d'abord, il permet d'obtenir plusieurs estimations des paramètres d'un modèle et non une seule à partir de la totalité des données. Dans le cas d'un modèle de viscosité, les paramètres pertinents ne sont a priori pas connus, comme le montre le grand nombre de modèles proposés dans la littérature. Cette approche permet de faire intervenir un grand nombre de paramètres et, par leur distribution, de juger de leur pertinence. Enfin, l'échantillonnage permet de diviser les données en une partition afin que le modèle construit à partir d'une partie des données (données de calibration) puisse être évalué sur des données qui n'ont pas servi à le construire (données de test).

3.2 Principe

Cette modélisation, qui a pour but de prédire la viscosité de fluides de gisement à partir des six variables d'entrée, se base sur les 196 points expérimentaux décrits précédemment. Cette base de données a été scindée en deux parties : 159 points de calibration ont servi pour les méthodes de régression à la construction des modèles statistiques ; les 37 points expérimentaux restants ont servi en tant que points de test pour l'évaluation des modèles obtenus. Une sélection entre les données de calibration et les données de test a alors été effectuée de telle sorte que les données correspondant aux frontières du domaine de validité servent à la calibration.

En outre, compte tenu de l'étendue des données de calibration comprenant des viscosités expérimentales allant de 0,21 à 10000 cP, il a été envisagé de déterminer non pas un modèle de prédiction mais trois, pour des classes de

viscosité différentes. Le recours à ce procédé nécessite alors la détermination d'une règle d'affectation par analyse factorielle discriminante [36, 37] à l'un des modèles pour prédire la viscosité de nouveaux fluides. Les classes de viscosité ont été créées de manière à réaliser un compromis entre :

– une bonne affectation par analyse factorielle discriminante (AFD),
– des modèles obtenus par régressions linéaire et non-linéaire pour les différentes classes.

Le modèle développé comprend alors trois classes de viscosité correspondant à trois modèles couvrant les gammes de viscosité suivantes : de 0 à 35 cP, de 35 à 500 cP et de 500 à 10000 cP. Le tableau 2 décrit la répartition des données utilisées pour la calibration et pour le test des trois modèles en fonction de leur classe de viscosité.

TABLEAU 2

Répartition des données dans chaque classe de viscosité
Distribution of data by class

Nombre de données	196	0 à 35 cP	35 à 500 cP	500 à 10000 cP
Données de calibration	159	89	46	24
Données de test	37	24	8	5

3.3 Modèles de régression

Nous disposons de deux méthodologies de modélisation basées l'une et l'autre sur la sélection de modèles par apprentissage.

La première méthodologie est basée sur l'association de la régression Partial Least Squares (*PLS*) avec le ré-échantillonnage par bootstrap (Aji *et al*., 2003, 2004; Efron et Tibshirani, 1993 ; Faraj *et al*., 2004, 2008 ; Lazraq *et al*., 2003 ; Tenenhaus, 1998) [38-44]. La deuxième concerne la régression non linéaire par réseaux de neurones combinée à un rééchantillonnage par validation croisée (K-folds) (Dreyfus *et al*., 2002 ; Hastie *et al*., 2001) [45, 46]. Dans la mesure du possible, pour des capacités prédictives équivalentes, le modèle de régression linéaire PLS est privilégié car ses paramètres sont directement liés aux variables d'entrées du modèle à la différence de la régression par réseaux de neurones. Le principe de ces méthodes est décrit en annexe.

3.3.1 Modèles de régression linéaire

La méthodologie, décrite ci-dessous, a été appliquée aux 3 classes de viscosité. Elle consiste à utiliser l'algorithme PLS-bootstrap sur les données d'une classe de viscosité en intégrant, en plus des effets principaux d'ordre 1, les effets principaux allant jusqu'à l'ordre 4 ainsi que les interactions d'ordre 2 et 3. Le nombre total de variables d'entrée est donc égal à 89. D'autre part, une transformation de BoxCox est systématiquement appliquée à la réponse afin d'en « normaliser » la distribution, ce qui améliore la qualité des modèles de prédiction. L'algorithme PLS-bootstrap permet de réduire le nombre de variables d'entrées en sélectionnant celles qui sont les plus pertinentes pour expliquer la viscosité.

Le nombre de bootstrap est fixé à 5000 et le seuil α à 0,15. Une fois que l'algorithme a convergé, la sélection de l'itération optimale s'effectue par l'étude des distributions des coefficients de validation (Q^2) pour chacune de ces itérations. L'itération optimale correspond alors à la distribution du Q^2 de variance minimale et de médiane maximale, les modèles évalués étant ceux affectés de la transformation inverse de BoxCox. Si l'itération optimale coïncide avec la dernière itération et que le nombre de variables pour cette itération est important, l'algorithme est relancé à partir des variables retenues pour cette itération mais le seuil α est diminué afin d'obtenir un critère de sélection plus stricte. Une fois l'itération optimale sélectionnée, 50 modèles PLS sont construits à partir d'échantillons bootstrap et des variables sélectionnées. Ces modèles sont ensuite appliqués aux données de test pour l'évaluation finale du modèle calculé.

Présentons maintenant les modèles de prédictions obtenus grâce à l'algorithme de PLS-Booststrap pour chacune des classes.

Sélection du nombre de paramètres du modèle à l'aide de l'algorithme PLS-bootstrap pour des valeurs comprises entre 0,21 et 35 cP.

Nous disposons d'un ensemble de données de calibration de 89 points expérimentaux. La figure 2 présente les résultats obtenus sur les points de viscosité non tirés par bootstrap (out-of-bag data) et permet de juger de la qualité de prédiction du modèle par validation croisée. On observe que 90 % des prédictions sont comprises dans un intervalle de confiance à ±10 % (ce dernier correspondant à l'incertitude sur les valeurs expérimentales). Au cours des itérations, les modèles obtenus sont donc de bonne qualité et l'itération optimale correspond à un modèle de 19 paramètres. Une fois le nombre de paramètres choisi, 50 modèles sont créés à partir de la totalité des données de cette classe de viscosité.

Sélection du nombre de paramètres du modèle à l'aide de l'algorithme PLS-bootstrap pour des valeurs comprises entre 35 et 500 cP.

La même démarche est appliquée aux données de la classe de viscosité comprises entre 35 et 500 cP. Le nombre de points de cette classe est égal à 46. La figure 3 présente les résultats obtenus sur les points de viscosité non tirés par bootstrap. Elle montre que les modèles de prédiction sont moins robustes que ceux déterminés pour la classe de viscosité de 0,21 à 35 cP. L'itération optimale correspond à un modèle de 13 paramètres.

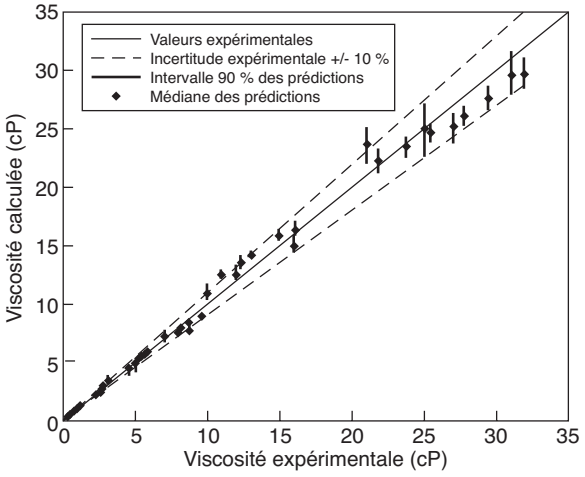

Figure 2

Graphique croisant les prédictions par validation croisée avec les valeurs de viscosité mesurées pour le modèle PLS-bootstrap pour la classe de viscosité de 0,21 à 35 cP.

Scatterplot of the predictions computed by crossvalidation with the values of viscosity measured for the PLS-bootstrap model for the viscosiy class from 0.21 to 35 cP.

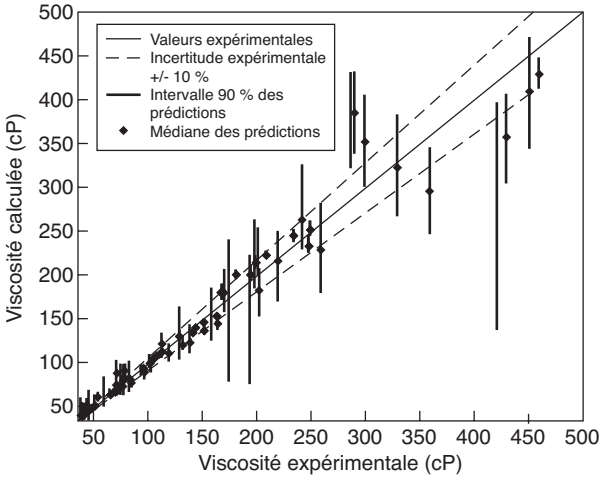

Figure 3

Graphique croisant les prédictions par validation croisée avec les valeurs de viscosité mesurées pour le modèle PLS-bootstrap pour la classe de viscosité de 35 à 500 cP.

Scatterplot of the predictions computed by crossvalidation with the values of viscosity measured for the PLS-bootstrap model for the viscosity class from 35 to 500 cP.

Une fois le nombre de paramètres choisi, 50 modèles sont créés à partir de la totalité des données de cette classe de viscosité.

Sélection du nombre de paramètres du modèle à l'aide de l'algorithme PLS-bootstrap pour des valeurs comprises entre 500 et 10000 cP.

La troisième classe contient 24 points. La figure 4 représente les résultats obtenus sur les points de viscosité non tirés par bootstrap. On constate que les distributions des prédictions de la plus grande partie des individus présentent un biais et une variabilité très importants. L'itération optimale correspond à un modèle de 14 paramètres. Une fois le nombre de paramètres choisi, 50 modèles sont créés à partir de la totalité des données de cette classe de viscosité.

Devant les résultats obtenus, nous observons que la méthodologie pour la modélisation de cette classe n'est pas adaptée aux fortes viscosités. Une part importante de cette difficulté peut sans doute être imputée au faible nombre de données disponibles pour cette classe. Il a donc été envisagé d'étudier une autre approche de régression pour cette classe.

3.3.2 Modèles de régression non linéaire

Les réseaux de neurones utilisés possèdent une couche de neurones cachés. La méthodologie utilisée permet de sélectionner le nombre optimal de neurones cachés pour la modélisation de la viscosité. Cette sélection est fondée sur l'utilisation d'une validation croisée (*K*-folds). Cette dernière permet d'évaluer le pouvoir prédictif des réseaux construits à partir des données non utilisées (*K* étant fixé à 5, le cinquième des données est utilisée pour la calibration). Cette méthode est expliquée en annexe B.

La modélisation à l'aide des réseaux de neurones de la classe contenant les viscosités comprises entre 0,21 et 35 cP n'est pas présentée car le modèle de régression linéaire construit a un pouvoir prédictif meilleur que celui obtenu par régression non linéaire.

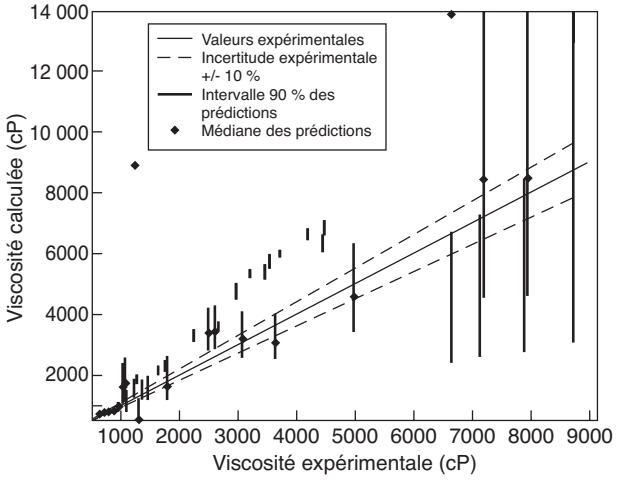

Figure 4

Graphique croisant les prédictions par validation croisée avec les valeurs de viscosité mesurées pour le modèle PLS-bootstrap pour la classe de viscosité de 500 à 10000 cP.

Scatterplot of the predictions computed by crossvalidation with the values of viscosity measured for the PLS-bootstrap mode for the viscosity class from 500 to 10000 cP.

Sélection du nombre de neurones à l'aide des réseaux de neurones pour des valeurs comprises entre 35 et 500 cP.
Les données de calibration étant au nombre de 46, il est envisagé de tester différentes architectures en faisant varier le nombre de neurones cachés. Le modèle obtenu correspond à une architecture à 2 neurones cachés. La figure 5 montre que les prédictions des modèles sur des données de calibration non utilisées pour créer le modèle sont peu biaisées mais présentent une variabilité importante.

Une fois l'architecture sélectionnée, 50 modèles sont créés à partir de la totalité des données de cette classe de viscosité.

Sélection du nombre de neurones à l'aide des réseaux de neurones pour des valeurs comprises entre 500 et 10000 cP.
Cette classe ne contient que 36 données, ce qui nous permet de comparer les modèles à 1, 2 ou 3 neurones. La figure 6 montre que les prédictions des modèles présentent une variabilité importante. Le modèle obtenu a une architecture à 2 neurones. Une fois l'architecture sélectionnée, 50 modèles sont créés à partir de la totalité des données de cette classe de viscosité.

3.4 Conclusions

Les modèles envisagés pour la prédiction de la viscosité des fluides de gisement sont donc :
- un modèle d'affectation de classe par analyse factorielle discriminante,
- de 0 à 35 cP : un modèle de régression linéaire PLS bootstrap,
- de 35 à 500 cP : pour cette classe de viscosité, le choix entre un modèle de régression linéaire ou non linéaire n'est pas évident. Aussi, l'application des deux modèles sur des données de test peut-elle nous orienter sur un choix de modèle, un modèle issu des réseaux de neurones,
- de 500 à 10000 cP : un modèle de réseaux de neurones.

3.5 Modèle d'affectation aux classes

La création de trois modèles de prédiction pour chacune des classes de viscosité nous a conduit à chercher un modèle d'affectation à partir des 159 points de calibration. La modélisation de l'affectation a été construite à partir de 50 modèles d'analyse factorielle discriminante établis sur 50 échantillons bootstrap. Outre la création de 50 modèles d'affectation, l'échantillonnage par bootstrap permet de déterminer que le nombre d'axes factoriels moyen optimal obtenu à partir des échantillons de calibration est égal à 5. Le calcul de la distance entre les projections des données et les 3 centres de classes dans l'espace engendré par les 5 axes factoriels permet alors d'estimer les probabilités d'appartenance des fluides à chacune des classes.

4 TEST DES MODÈLES ÉTUDIÉS

Cette dernière partie permet de valider les modèles précédemment déterminés sur les points de test (*i.e.* ceux ayant été écartés de la phase de construction de ces modèles). Cette

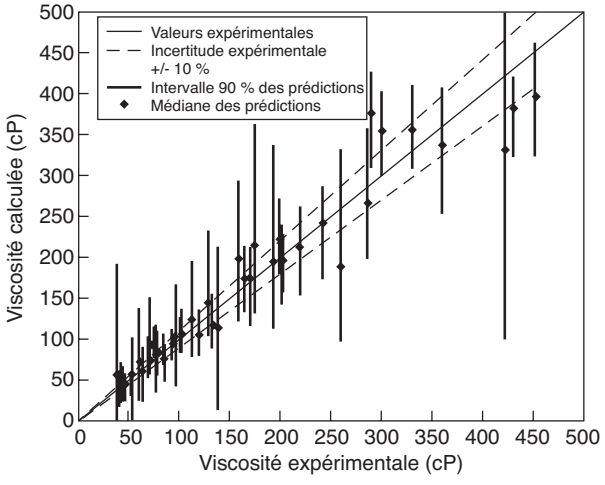

Figure 5

Graphique croisant les prédictions par validation croisée avec les valeurs de viscosité mesurées pour le modèle de réseaux à 2 neurones et la classe de viscosité de 35 à 500 cP.

Scatterplot of the predictions computed by crossvalidation with the values of viscosity measured for the Neural Networks with 2 hidden units for viscosity class from 35 to 500 cP.

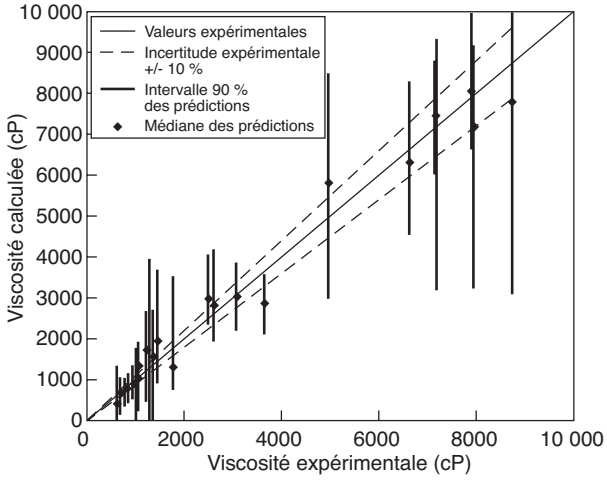

Figure 6

Graphique croisant les prédictions par validation croisée des fluides de la troisième classe avec les valeurs de viscosité mesurées pour le modèle de réseaux à 2 neurones.

Scatterplot of the prediction with the values of viscosity measured for the Neural Networks with 2 hidden units for viscosity class from 500 cp to 10000 cP.

validation est basée sur le critère de l'écart relatif défini de la manière suivante :

$$\text{Écart relatif}\,(\%) = 100 \times \frac{\left|\text{viscosité}_{\text{calc}} - \text{viscosité}_{\text{exp}}\right|}{\text{viscosité}_{\text{exp}}}$$

4.1 Test du modèle d'affectation aux classes

L'étude de la fiabilité du modèle d'affectation aux classes a tout d'abord été effectuée. Les résultats montrent alors que le pourcentage des fluides bien classés, calculé à partir de la moyenne des probabilités d'appartenance aux 3 classes, est égal à 100 % pour chacune des classes. Nous disposons ainsi d'un outil fiable d'affectation de classe permettant d'utiliser pour chaque fluide étudié le bon modèle de prédiction.

4.2 Test des modèles de prédiction

Les principaux outils utilisés pour évaluer ces modèles sont, l'écart relatif détaillé dans l'équation précédente et la variabilité de la prédiction (*i.e.* distribution des prédictions) matérialisée par les lignes verticales sur les figures 7 et 8. Les résultats sont les suivants.

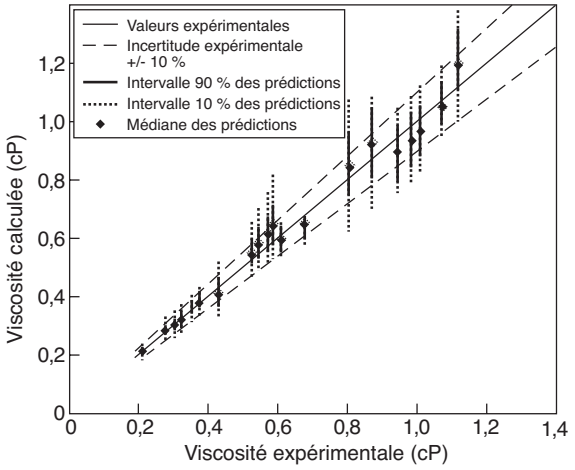

Figure 7

Prédiction des viscosités pour la première classe de viscosité jusqu'à 1,4 cP.

Predicted values of viscosities for 0 to 1.4 cP.

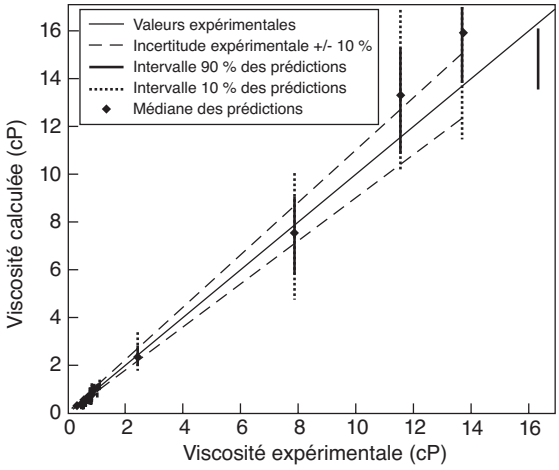

Figure 8

Prédiction des viscosités pour la première classe de viscosité jusqu'à 16 cP.

Predicted values of viscosities for 0 to 16 cP.

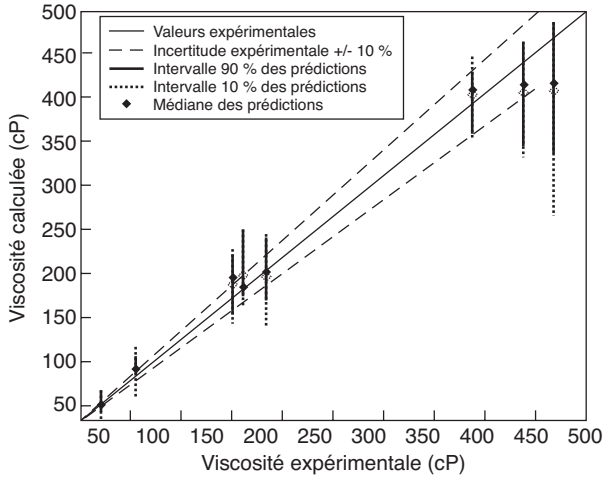

Figure 9

Prédiction des viscosités pour la classe 35 à 500 cP.

Predicted values of viscosities for 35 to 500 cP.

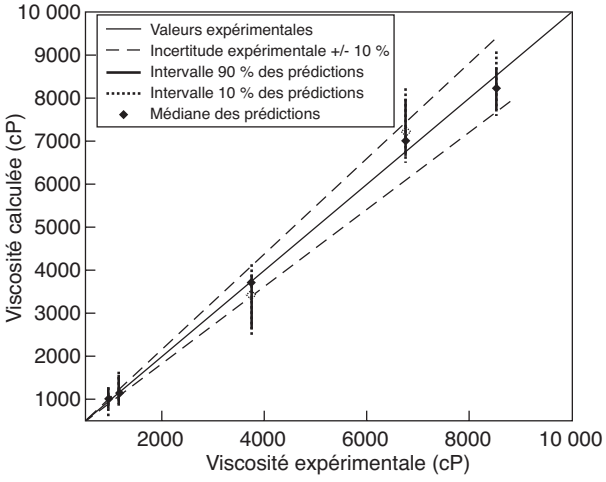

Figure 10

Prédiction des viscosités pour la classe 500 à 10000 cP.

Predicted values of viscosities for 500 to 10000 cP.

Modèle issu de l'algorithme PLS-bootstrap pour des valeurs comprises entre 0,21 et 35 cP.

L'écart relatif moyen obtenu pour cette classe de viscosité est de 7,96 %. De plus, les figures 7 et 8 présentent de bonnes variabilités autour des points étudiés. La majeure partie des points est prédite dans l'intervalle de confiance de + ou –10 % autour de l'idéalité.

Modèle issu des réseaux de Neurones pour des valeurs comprises entre 35 et 500 cP.

En ce qui concerne cette classe de viscosité, les deux modèles de régression linéaire et non linéaire ont été appliqués aux données de test. La figure 9 présente les résultats obtenus par le modèle neuronal. En ce qui concerne ces données de test, ce modèle est plus performant que la régression PLS. Ce résultat devrait cependant être considéré avec prudence compte tenu du nombre peu élevé des points de test.

L'écart relatif moyen pour cette classe de viscosité est de 6,47 % et la variabilité des points étudiés sur la figure 9 semble raisonnable.

Modèle issu des réseaux de Neurones pour des valeurs comprises entre 500 et 10000 cP.

Pour cette classe de bruts lourds, l'écart relatif moyen est de 4,26 % et la distribution des prédictions autour de chaque point sur la figure 10 est bonne.

CONCLUSION

L'approche statistique présentée pour la modélisation de la viscosité des fluides de gisement par les méthodes de régression linéaire et non-linéaire permet d'obtenir de bons résultats sur un large éventail de viscosités s'échelonnant de 0 à 10000 cP. Cette méthode de prédiction reste cependant améliorable compte tenu des incertitudes liées aux mesures de la viscosité des fluides de gisement, des conditions de pression et de température variées qu'ils subissent au sein du réservoir et de la composition unique qui les caractérise. Cette analyse va servir à la préparation d'une prochaine étude dans laquelle les données expérimentales seront mises en place suivant un plan d'expériences permettant de recouvrir l'ensemble du domaine des variables en entrée du modèle.

RÉFÉRENCES

1 Beggs H.D., Robinson J.R. (1975) Estimating the Viscosity of Crude Oil Systems, *J. Petrol. Technol.* **9**, 1140-1141.

2 Petrosky G.E., Farshad F.F. (1995) Viscosity Correlations for Gulf of Mexico Crudes Oils, *SPE, 29468.*

3 Labedi R. (1992) Improved Correlations for Predicting the Viscosity of Light Crudes, *J. Petrol. Sci. Eng.* **8**, 221-234.

4 Khan S.A., Al-Marhoun M.A., Duffuaa S.O., Abu-Khamsin S.A. (1987) Viscosity Correlations for Saudi Arabian Crude Oils, *SPE,* 15 720.

5 Andrade E.N. (1934), *Philos. Mag.* **17**, 497.

6 Tait P.G. (1888) The Voyage of H.M.S. Challenger, Report on some of the Physical Properties of Fresh Water and of Sea-Water, *Phys. Chem. Chall. Exp*, part 4.

7 Dymond J. H., Oye H. A. (1994) Viscosity of Selected Liquid Hydrocarbons, *J. Phys. Chem. Ref. Data* **23**, 1, 41-53.

8 Tanaka Y., Hosokawa H., Kubota H., Makita T. (1991) Viscosity and Density of Binary Mixtures of Cyclohexane with n-Octane, n-Dodecane, and n-Hexadecane under High Pressures, *Int. J. Thermophys.* **12**, 2, 245-264.

9 Assael M J., Karagiannidis L., Papadaki M. (1991) Measurements of the Viscosity of n-Heptane + n-Undecane Mixtures at Pressures up to 75 Mpa, *Int. J. Thermophys.* **12**, 5, 811-820.

10 Comunas M.J.P., Baylaucq A., Boned C., Fernandez J., (2001) High Pressure Measurements of the Viscosity and Density of Two Polyethers and Two Dialkyl Carbonates, *Int. J. Thermophys.* **22**, 3, 749-768.

11 Kanti M., Zhou H., Ye S., Boned C., Lagourette B., Saint-Guirons H., Xans P., Montel F. (1989) Viscosity of Liquid Hydrocarbons, Mixtures and Petroleum Cuts, as a Function of Pressure and Temperature, *J. Phys. Chem.* **93**, 3860-3864.

12 Van der Waals J.D., *PhD Thesis*, 1873, Leyden.

13 Pedersen K.S., Fredenslund A. (1984) Viscosity of Crude Oil, *Chem. Eng. Sci.* **39**, 6, 1011-1016.

14 Lohrenz J., Bray B.G., Clark C.R. (1964) Calculating Viscosities of Reservoir Fluids from their Composition, *J. Petrol. Technol.* 1171-1176.

15 Little J.E., Kennedy H.T. (1968) A Correlation of the Viscosity of Hydrocarbon Systems with Pressure, Temperature and Composition, *Soc. Petrol. Eng. J. AIME*, **243**, 157-162.

16 Guo X.Q., Wang S.X., Rong S.X., Guo T.M.(1997) Viscosity Model Based on Equations of State for Hydrocarbon Liquids and Gases (1997) *Fluid Phase Equilib.* **139**, 405-421.

17 Patel N.C., Teja A.S. (1982) A New Cubic Equation of State for Fluids and Fluid Mixtures, *Chem. Eng. Sci.* **37**, 463-473.

18 Vidal J. (2003) chapter 4, in *Thermodynamics: Applications in Chemical Engineering and the Petroleum Industry, 2003*, Ed. Technip, Paris.

19 Quiñones-Cisneros S.E., Zeberg-Mikkelsen C.K., Stenby E.H. (2001) Friction Theory for Viscosity Modeling: Extension to Crude Oil, *Chem. Eng. Sci.* **56**, 7007-7015.

20 Quiñones-Cisneros S.E., Zeberg-Mikkelsen C.K., Baylaucq A., Boned C. (2004) Viscosity Modeling and Prediction of Reservoir Fluids: from Natural Gas to Heavy Oils, *Int. J. Thermophys.* **25**, 5, 1353-1366.

21 Cohen M.H., Turnbull D. (1959) Molecular Transport in Liquids and Glasses, *J. Chem. Phys.* **31**, 5, 1164-1169.

22 Doolittle A.K. (1951) Studies in Newtonian Flow. II. The Dependence of the Viscosity of Liquids on Free-Space, *J. Appl. Phys.* **22**, 12, 1471-1475.

23 Allal A., Boned C., Baylaucq A. (2001) Free Volume Viscosity Model for Fluids in the Dense and Gaseous States, *Phys. Rev. E* **64**, 011203.

24 Boned C., Allal A., Baylaucq A., Zeberg-Mikkelsen C.K., Bessières D., Quinones-Cisneros S.E. (2004) Simultaneous Free-volume Modeling of the Self-Diffusion Coefficient and Dynamic Viscosity at High Pressure, *Phys. Rev. E* **69**, 031203.

25 Elsharkawy A.M., Hassan S.A., Hasim Y.S.K., Fahim M.A. (2003) New Compositional Models for Calculating the Viscosity of Crude Oils, *Ind. Eng. Chem. Res.* **42**, 4132-4142.

26 Werner A., De Hemptinne J.C., Behar F., Behar E., Boned C. (1998) A New Viscosity Model for Petroleum Fluids with High Asphaltenes Content, *Fluid Phase Equil.* **147**, 319-341.

27 Werner A., Behar F., De Hemptinne J.C., Behar E. (1998) Viscosity and Phase Behaviour of Petroleum Fluids with High Asphaltenes Content, *Fluid Phase Equil.* **147**, 343-356.

28 Werner A. (1996) Viscosité des fluides pétroliers riches en produits lourds Mesure et modélisation, *Thèse présentée à l'Université de Pau et des pays de l'Adour, France.*

29 Grunberg L., Nissan, A.H. (1949) Mixture Law for Viscosity, *Nature* **164**, 4175, 799-800.

30 Hubbard R.M., Brown G.G. (1943) The Rolling Ball Viscometer, *Ind. Eng. Chem.* **15**, 3, 212-219.

31 Stanislawki U., Luft G. (1987) Dynamic Viscosity of Ethene, *Ber. Bunsenges Phys. Chem.* **91**, 756-759.

32 Sawamura S., Takeuchi N., Kitamura K., Taniguchi Y. (1990) High Pressure Rolling Ball Viscometer of a Corrosion-Resistant Type, *Rev. Sci. Instrum.* **61**, 2, 871-873.

33 Barrufet M.A., Hall K.R., Estrada-Baltazar A., Iglesias-Silva G.A. (1999) Liquid Viscosiy of Octane and Pentane + Octane Mixtures from 298.15 K to 373.15 K up to 25 MPa, *J. Chem. Eng. Data* **44**, 1310-1314.

34 Pensado A.S., Communas M.J.M., Lugo L., Fernandez J. (2005) Experimental Dynamic Viscosities of 2,3-Dimethylpentane up to 60 MPa and from (303.15 to 353.15) K using a Rolling Ball Viscometer, *J. Chem. Eng. Data* **50**, 849-855.

35 Arnaud J.F. (1995) Caractérisation des propriétés physiques et thermodynamiques des fluides pétroliers à hautes pressions, *Thèse présentée à l'Université de Pau et des pays de l'Adour, France.*

36 Saporta G. (2006) *Probabilités analyse des données et statistique*, 2nd éd., Technip, 442 (a), 493 (b).

37 Lebart L., Morineau A., Piron M. (1995) *Statistique exploratoire multidimensionnelle*, Dunod, 251-282 (a), 347-370 (b).

38 Aji S., Tavolaro S., Lantz F., Faraj A. (2003) Apport du bootstrap à la régression PLS : application à la prédiction de la qualité des gazoles, *Oil Gas Sci. Technol. - Rev. IFP* **58**, 5, 599-608.

38 Aji S., Shildknecht-Szydlowski N., Faraj A. (2004) Partial Least Square Modeling for the Control of Refining Processes on Mid-Distillates by Near Infrared Spectroscopy, *Oil Gas Sci. Technol. - Rev. IFP* **59**, 3, 303-321.

40 Efron B., Tibshirani R. (1993) *An Introduction to the Bootstrap*, Chapman and Hall, London.

41 Faraj, A., Constant, M. (2004) Utilisation du bootstrap pour la sélection de variables et la typologie des individus en Régression PLS, *actes des 36es journées de Statistique de la SFDS,* Montpellier, 24 au 28 mai 2004.

42 Faraj A., Noçairi H., Constant M. (2008) Sélection de modèle PLS par rééchantillonnage bootstrap : points de vue croisés, *Numéro spécial de la Revue des Nouvelles Technologies de l'Information*, avril 2008, 185-200.

43 Lazraq A., Cléroux R., Gauchi J.P. (2003) Selecting both Latent and Explanatory Variables in the PLS1 Regression Model, *Chemometrics and Intelligent Laboratory Systems* **66**, 117-126.

44 Tenenhaus, M. (1998) *La Régression PLS Théorie et Pratique*, Technip.

45 Dreyfus G., Martinez J.M., Samuelides M., Gordon M.B., Badran F., Thiria S., Hérault L. (2002) *Réseaux de neurones Méthodologie et applications*, Eyrolles.

46 Hastie T., Tibshirani R., Fiedman J. (2001) *The Elements of Statistical Learning*, Springer Series in Statistics, 347-370.

Manuscrit final reçu en janvier 2008
Publié en ligne en septembre 2008

Copyright © 2008 Institut français du pétrole

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than IFP must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee: Request permission from Documentation, Institut français du pétrole, fax. +33 1 47 52 70 78, or revueogst@ifp.fr.

ANNEXES

A. Algorithme PLS-bootstrap

Voici les différentes étapes de l'algorithme. Un certain nombre d'indices permettant de juger de la qualité des modèles sont présentés en plus de ceux utilisés dans l'étude. La régression PLS est décrite de manière détaillée dans le livre de Tenenhaus (1998) [44]. En ce qui concerne la méthodologie PLS-bootstrap, il est conseillé de se référer aux articles de Aji *et al*. (2003) [38], Aji *et al*. (2004) [39], Faraj *et al*. (2004) [41], Faraj *et al*. (2008) [42], Lazraq *el al*. (2003) [43].

Étape 1 : Répéter pour ℓ = 1, 2, ..., L

1 Construire un échantillon aléatoire $\mathbf{Z}^{*\ell}$ de taille N tiré avec remise dans Z : $\mathbf{Z}^{*\ell} = \{(x_i, y_i), i \in C^{*\ell}\}$ est l'échantillon bootstrap ℓ.

$C^{*\ell}$ est l'ensemble des indices des individus ayant été tirés (certains peuvent être dupliqués plusieurs fois ; on a $N = |C^{*\ell}|$). Cet ensemble est appelé ensemble d'apprentissage.

2 Construire l'échantillon des individus non tirés : $\overline{\mathbf{Z}}^{*\ell} = \{(x_i, y_i), i \in \overline{C}^{*\ell}$. cet ensemble correspond aux données de validation croisée (out-of-bag-data).

$\overline{C}^{*\ell}$ est l'ensemble des indices des individus n'ayant pas été tirés dans l'échantillon bootstrap ℓ.

3 Construire le modèle $\hat{\mathbf{y}}^{*\ell} = \mathbf{X}^{*\ell}\mathbf{b}^{*\ell}$ de Régression PLS où $\mathbf{b}^{*\ell} = (b_1^{*\ell}, b_2^{*\ell}, \ldots, b_j^{*\ell})^T$ est le vecteur colonne des coefficients du modèle.

4 Calculer les prédictions du modèle pour les individus i non tirés (*i.e.* $i \in \overline{C}^{*\ell}$) : $\hat{y}_i^{*\ell} = (\mathbf{b}^{*\ell})^T \mathbf{x}_i^{*\ell}$ où le vecteur $\mathbf{x}_i^{*\ell} = (\mathbf{x}^1{}_i^{*\ell}, \mathbf{x}^2{}_i^{*\ell}, \ldots, \mathbf{x}^J{}_i^{*\ell})T$ représente la $i$ème ligne de la matrice = $\mathbf{X}^{*\ell}(\mathbf{x}_i)$ où $i \in C^{*\ell}$.

5 Calculer le coefficient de détermination $R^{*\ell 2}$ à partir des données de calibration et le $Q^{*\ell 2}$ correspondant au coefficient de validation calculé à partir des données de validation croisée (out-of-bag-data).

Étape 2 : Calculer pour i = 1, ..., N

6 Les ensembles $\Lambda^i = \{\ell, i \in C^{*\ell}\}$ des indices ℓ des échantillons bootstrap contenant i et $\Lambda^i = \{\ell, i \in \overline{C}^{*\ell}\}$ des indices ℓ des échantillons bootstrap ne contenant pas i. On note respectivement leur cardinal $|\Lambda^i|$ et $|\Lambda^{-i}|$.

7 La variance de prédiction σ_i^{*2} à partir des $|\Lambda^{-i}|$ modèles bootstrap.

8 Le biais B_i^* de prédiction à partir des $|\Lambda^{-i}|$ modèles bootstrap.

9 L'erreur de prédiction $e_i^{*\ell}$ pour chaque bootstrap $\ell, \in \Lambda^{-i}$.

Étape 3 : Répéter pour j = 1, 2, ..., J

10 Calculer l'intervalle de confiance, à un seuil α fixé, $I^*j(\alpha)$ pour le coefficient b_j de la variable x_j à partir de l'échantillon bootstrap $E_j = \{b_j^{*\ell}, \ell = 1, L\}$

11 Si $0 \in I^*j\ (\alpha)$, éliminer la variable x_j.

Répéter les étapes 1 à 3 avec les variables X^j retenues, jusqu'à ce qu'aucune variable ne soit éliminée.

Les valeurs de $R^{*\ell 2}$, $Q^{*\ell 2}$, σ_i^{*2}, B_i^* et $e_i^{*\ell}$, calculées lors des étapes 1 et 2, sont définies de la façon suivante :

$$R^{*\ell 2} = cor^2(\hat{\mathbf{y}}^{*\ell}, \mathbf{y}^{*\ell})$$

$$Q^{*\ell 2} = cor^2(\hat{\mathbf{y}}^{*-\ell}, \mathbf{y}^{*-\ell})$$

$$\sigma_i^{*2} = \frac{1}{\left|C^{-i}\right|} \sum_{\ell \in C^{-i}} (\hat{y}_i^{*\ell} - \overline{y}_i^*)^2$$

$$B_i^* = \left| y_i - \overline{y}_i^* \right|$$

$$e_i^{*\ell} = \hat{y}_i^{*\ell} - y_i \text{ pour } \ell \in \Lambda^{-i}$$

où :

$$\overline{y}^{*\ell} = \frac{1}{\left|C^{*\ell}\right|} \sum_{\ell \in C^{*\ell}} y_i$$

est la moyenne de y calculée sur l'échantillon bootstrap $C^{*\ell}$.

$$\overline{y}_i^* = \frac{1}{\left|\Lambda^{-i}\right|} \sum_{\ell \in \Lambda^{-i}} \hat{y}_i^{*\ell}$$

est la moyenne des prédictions des L modèles bootstrap au point i.

$\Lambda^{-i} = \{\ell, i \in \overline{C}^{*\ell}\}$ est l'ensemble des indices des échantillon bootstrap ℓ ne contenant pas i.

Son cardinal est noté : $|\Lambda^{-i}|$= card (Λ^{-i}) et $\overline{C}^{*\ell}$ est l'ensemble des indices des individus n'ayant pas été tirés dans l'échantillon bootstrap ℓ.

σ_i^{*2} est une estimation de la variance de prédiction du modèle de Régression PLS pour l'individu i. C'est un indicateur de la dispersion de la distribution empirique $\{\hat{y}_i^{*\ell}\}$ de la prédiction du modèle au point i. Grâce à σ_i^{*2} on peut estimer la variance de prédiction en tout point du domaine de variation des variables explicatives.

B_i mesure la justesse du modèle – écart absolu entre la valeur mesurée y_i et la prédiction moyenne du modèle – au point i.

Les valeurs de $Q^{*\ell 2}$, $\hat{y}_i^{*\ell}$, B_i^*, $e_i^{*\ell}$ et σ_i^{*2}, bien que ne servant pas directement dans l'algorithme de sélection des variables, aident à représenter les liens existant entre les variables, entre les individus et entre les individus et les variables. Les graphiques de ces valeurs rendent compte de la qualité des modèles construits et renseignent sur la nature (linéaire ou non linéaire) de ces modèles. Ils permettent, de cette façon, de distinguer la (ou les) itération(s) correspondant

aux meilleurs ensembles de variables sélectionnées (*i.e.* celles associées aux modèles dont les qualités de prédiction sont les meilleures).

B. Réseaux de neurones

Les Réseaux de Neurones utilisés dans cette étude sont de type Perceptrons MultiCouches (*MLP* pour Multi Layer Perceptron) à une couche de neurones cachés [45, 46].

Soit $\mathbf{X} = (\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^J)$ la matrice $J \times N$ des J variables explicatives ($\mathbf{x}^j$ est le vecteur de dimension N représentant la j-ième colonne de $\mathbf{X}$) et y la matrice de la variable à expliquer. $\mathbf{X}$ et y forment les données d'apprentissage car c'est à partir de ces données que le modèle de prédiction est construit. Voici la forme de ce modèle :

$$g(\mathbf{X}, \boldsymbol{\theta}) = b_0 + \sum_{q=1}^{Q} w_q \tanh\Big(\sum_{j=1}^{J} \big(b_q + w_{jq}\mathbf{x}^j\big)\Big)$$

$\boldsymbol{\theta}$ contenant les paramètres du réseau (w et b), Q étant le nombre de neurones de la couche cachée et J le nombre de variables en entrée.

Cette équation indique que le nombre paramètres du réseau dépend du nombre de variables d'entrées et du nombre de neurones dans la couche cachée. Afin de pouvoir construire le réseau, il est nécessaire que le nombre de données d'apprentissage soit, au minimum, égal au nombre de paramètres du réseau.

D'autre part, cette fonction a la propriété d'être un approximateur universel : c'est-à-dire qu'avec un nombre fini de neurones, elle peut modéliser n'importe quelle fonction continue bornée. De manière pratique, cela signifie que les réseaux de neurones peuvent modéliser presque parfaitement la réponse correspondant aux données d'apprentissage à partir du moment où le nombre de ces données est suffisant. Le risque est, cependant, de créer un modèle dont le pouvoir de généralisation à de nouvelles prédictions soit faible. Ce risque est directement lié à la structure du réseau. En effet, plus le réseau contient de neurones, plus il permet de modéliser des phénomènes complexes et propres aux données d'apprentissages. Il s'agit donc de trouver un modèle dont le pouvoir de généralisation est important En modélisation statistique, l'objectif n'est pas de créer un modèle parfaitement ajusté aux données d'apprentissage mais plutôt de trouver un modèle dont les prédictions peuvent être étendues à de nouvelles données expérimentales. C'est dans cette optique qu'il a été choisi d'associer à la modélisation neuronale une méthode d'échantillonnage par validation croisée (K-folds).

Le principe de cette méthodologie est de déterminer le nombre de neurones optimal du point de vue de la généralisation à de nouvelles données. La validation croisée (K-folds) est une méthode d'échantillonnage consistant à construire, à partir d'un échantillon de calibration, des échantillons d'apprentissages (utilisés pour la construction de modèles) et de validation (utilisés pour l'évaluation ou la comparaison des modèles). Voici la manière dont ces derniers sont déterminés.

Les N individus appartenant à l'échantillon de calibration sont divisés en K groupes de manière aléatoire. Le modèle est construit à partir de $K - 1$ groupes formant l'échantillon d'apprentissage, l'échantillon restant formant l'échantillon de validation. Cette opération est réitérée K fois, en écartant à chaque fois un groupe différent pour la validation. Chaque individu participe donc $K - 1$ fois à la construction de $K - 1$ modèles et une fois à l'évaluation d'un modèle construit sans sa participation.

Il s'agit donc de construire et de comparer H modèles neuronaux qui diffèrent par leur nombre de neurones. La comparaison de leurs performances prédictives se fait grâce aux échantillons de validation. Le modèle sélectionné est bien évidemment celui dont l'erreur de prédiction est la plus faible.

Afin d'obtenir, pour une architecture h donnée, une distribution du coefficient de validation (Q^2), I initialisations des poids (paramètres) du réseau sont effectuées. Les I modèles créés, de cette manière, permettent d'obtenir I prédictions des fluides en validation et donc d'obtenir une distribution du Q^2.

Au total $H{\cdot}I{\cdot}(K - 1)$ modèles d'apprentissages sont déterminés afin de sélectionner l'architecture optimale. Une fois que celle-ci est déterminée, l'ensemble des données de calibration est utilisé pour la détermination de I' modèles de calibration constitués à partir de I' initialisations du réseau avec l'architecture.

C. Modèles déterminés par PLS-bootstrap

Ils ont la forme suivante :

$$\mathbf{y} = (\lambda \mathbf{y}_{\mathbf{transf}} + 1)^{\frac{1}{\lambda}}$$

$$\text{avec } \mathbf{y}_{\mathbf{transf}} = \mathbf{X}_{\beta} \cdot \beta$$

où $\mathbf{X}_{\beta}$ désigne la matrice contenant les variables correspondant aux coefficients du modèle contenus dans β.
λ est le coefficient de transformation utilisé.

- *Modèle de viscosité pour des viscosités inférieures à 35 cP*

Le tableau ci-après donne les 19 entrées du modèle ainsi que la constante du modèle et les coefficients moyens leur correspondant.

avec :

T la température exprimée en °C,
P la pression exprimée en bar et la viscosité en cP,
*W*1 la fraction massique Gaz W_{gaz} (hydrocarbures C_1-C_5 et non hydrocarbures),
*W*2 la fraction massique $W_{C6\text{-}C20}$,
*W*3 la fraction massique $W_{C20+\ S\text{-}A\text{-}R}$ (Saturés, Aromatiques, Résines),
*W*4 la fraction massique $W_{C20+\ asph}$ (asphaltènes).

	Variable	Coefficient moyen
1	$P \times W4^2$	6,88E-06
2	$P \times W2^2$	-1,03E-06
3	$T \times W4^2$	-1,25E-04
4	$T \times W1^2$	-4,82E-05
5	$W4 \times T^2$	1,17E-05
6	$W4 \times W1^2$	8,36E-04
7	$W1 \times W3^2$	-4,41E-05
8	$W2 \times W4 \times P$	-3,88E-06
9	$W2 \times W4 \times T$	-6,78E-05
10	$W2 \times W3 \times T$	-1,16E-05
11	$W2 \times W3 \times W4$	4,80E-05
12	$W1 \times T \times P$	2,03E-07
13	$W1 \times W4 \times T$	-4,52E-05
14	$W1 \times W2 \times T$	1,88E-05
15	$W2 \times P$	6,85E-05
16	$W1^4$	-3,55E-06
17	T^2	3,00E-05
18	$W3^2$	5,91E-04
19	$W4$	0,1776
20	Constante	0,9182

Le coefficient de transformation BoxCox pour ce modèle est –0,22.

- *Modèle de viscosité pour des viscosités allant de 35 cP à 500 cP*

Le tableau ci-dessous donne les 13 entrées du modèle ainsi que la constante du modèle et les coefficients moyens leur correspondant.

	Variable	Coefficient moyen
1	$T \times W3^2$	-6,31E-06
2	$W3 \times W1^2$	1,35E-04
3	$W1 \times W4^2$	-2,38E-03
4	$W1 \times W2^2$	-2,15E-03
5	$W3 \times T \times P$	-1,31E-07
6	$W2 \times W3 \times W4$	-8,27E-05
7	$W1 \times W4 \times P$	-2,57E-05
8	$W1 \times W2 \times W4$	3,85E-0
9	$W4^4$	2,15E-05
10	$W2^4$	2,11E-05
11	P^3	3,97E-09
12	$W2^3$	-3,39E-0
13	$W2^2$	8,86E-02
14	Constante	-3,2703

avec :

T la température exprimée en °C,

P la pression exprimée en bar et la viscosité en cP,

$W1$ la fraction massique Gaz W_{gaz} (hydrocarbures C_1-C_5 et non hydrocarbures),

$W2$ la fraction massique $W_{C6\text{-}C20}$,

$W3$ la fraction massique $W_{C20+\ S\text{-}A\text{-}R}$ (Saturés, Aromatiques, Résines),

$W4$ la fraction massique $W_{C20+\ asph}$ (asphaltènes).

Le coefficient de transformation BoxCox pour ce modèle est –0,12.

- *Modèle de viscosité pour des viscosités allant de 500 cP à 10000 cP*

Le tableau ci-dessous donne les 14 entrées du modèle ainsi que la constante du modèle du modèle et les coefficients moyens leur correspondant.

	Variable	Coefficient moyen
1	$T \times W4^2$	-5,61E-05
2	$W4 \times T^2$	-5,12E-07
3	$W4 \times W3^2$	5,13E-06
4	$W4 \times W2^2$	1,62E-05
5	$W3 \times W4 \times T$	-2,97E-06
6	$W2 \times W4 \times T$	1,45E-05
7	$W1 \times W3 \times T$	-2,27E-05
8	$W4 \times T$	-1,92E-04
9	$W3 \times W4$	7,82E-04
10	$W1 \times T$	-1,43E-04
11	$W2^4$	-8,37E-07
12	$W2^3$	-8,20E-06
13	P^2	3,41E-06
14	$W2^2$	2,79E-04
15	Constante	2,7531

avec

T la température exprimée en °C,

P la pression exprimée en bar et la viscosité en cP,

$W1$ la fraction massique Gaz W_{gaz} (hydrocarbures C_1-C_5 et non hydrocarbures),

$W2$ la fraction massique $W_{C6\text{-}C20}$,

$W3$ la fraction massique $W_{C20+\ S\text{-}A\text{-}R}$ (Saturés, Aromatiques, Résines),

$W4$ la fraction massique $W_{C20+\ asph}$ (asphaltènes).

Le coefficient de transformation BoxCox pour ce modèle est –0,28.

D. MODÈLES DÉTERMINÉS PAR RÉSEAUX DE NEURONES

Les modèles créés par le réseau de neurones sont de la forme :

$$g(\mathbf{X},\boldsymbol{\theta}) = b_0 + \sum_{q=1}^{Q} w_q \tanh\left(\sum_{j=1}^{J} (b_q + w_{jq}\mathbf{x}_j)\right)$$

$\boldsymbol{\theta}$ est le vecteur des paramètres poids w et b.

où : $\mathbf{X}$ est un nouvel individu définit par les variables

x_1 son pourcentage massique en gaz,

x_2 son pourcentage massique en C6-C20,

x_3 son pourcentage massique en SAR,

x_4 son pourcentage massique en asphaltène,

x_5 sa température en °C,

x_6 sa pression en bar.

- *Modèle de viscosité pour des viscosités allant de 35 cP à 500 cP*

Les tableaux ci-après présentent des coefficients obtenus pour ce modèle.

avec w_{jq} :

0,2032	0,6500
0,4658	0,1890
-0,0372	-0,0785
0,7826	-0,2559
0,0451	0,0578
0,0069	-0,0028

b_q :

-18,6219	0,3110

w_q :

-155,7450	-340,6276

b_0 :

535,4017

- *Modèle de viscosité pour des viscosités allant de 500 cP à 10000 cP*

Les tableaux ci-dessous présentent des coefficients obtenus pour ce modèle.

avec w_{jq} :

-3,9518	3,656
0,3269	0,3320
-0,1274	-0,1570
0,3456	-0,2746
0,1348	0,1452
0,0255	-0,0082

b_q :

-9,0727	2,1527

w_q :

-0,8171	-4,8541

b_0 :

6,55E+03