Reconstruction of Petroleum Feedstocks by Entropy Maximization. Application to FCC Gasolines

D. Hudebine, J.J. Verstraete

Oil & Gas Science and Technology – Rev. IFP Energies nouvelles, Vol. 66 (2011), No. 3, pp. 437-460

DOI: 10.2516/ogst/2011110

Dossier

Chemical Reaction Modelling of Refining Processes
Modélisation cinétique des procédés de raffinage

Reconstruction of Petroleum Feedstocks by Entropy Maximization. Application to FCC Gasolines

D. Hudebine* and J.J. Verstraete

IFP Energies nouvelles, Rond-point de l'échangeur de Solaize, BP 3, 69360 Solaize - France
e-mail: damien.hudebine@ifpen.fr - jan.verstraete@ifpen.fr

Résumé — Reconstruction de coupes pétrolières par maximisation d'entropie. Application aux essences de FCC — Dans le domaine pétrolier, les coupes sont généralement des mélanges complexes de plusieurs centaines à plusieurs millions d'espèces chimiques différentes. De ce fait, les outils analytiques, même les plus performants, ne permettent pas de séparer et d'identifier l'ensemble des composés présents. Les fractions pétrolières sont donc actuellement caractérisées soit via des descripteurs macroscopiques moyens (densité, analyse élémentaire, résonance magnétique nucléaire, etc.), soit à l'aide de techniques séparatives (distillation, chromatographie gaz ou liquide, spectrométrie de masse, etc.) qui ne quantifient cependant que quelques grandes familles de molécules. Les méthodes de reconstruction de coupes pétrolières sont des approches informatiques qui permettent d'évoluer vers un détail plus moléculaire en se basant sur le principe suivant : définir des mélanges simplifiés mais cohérents de composés chimiques à partir de données analytiques parcellaires et de connaissances expertes du procédé étudié. Ainsi, la méthode de reconstruction par maximisation d'entropie, proposée dans cet article, est une technique récente et puissante permettant de déterminer les fractions molaires d'un mélange préétabli de composés chimiques en maximisant un critère entropique et en respectant les contraintes analytiques fixées par le modélisateur. L'utilisation de cette méthodologie permet de réduire le nombre de degrés de liberté du système de quelques milliers (correspondant aux fractions molaires des composés) à quelques dizaines (correspondant aux paramètres de Lagrange associés aux contraintes analytiques) et ainsi de diminuer fortement le temps de calcul nécessaire à la résolution du problème. Cette approche a été appliquée avec succès à la reconstruction d'essences de FCC en prédisant précisément la composition moléculaire de ce type de coupes pétrolières à partir d'une distillation simulée et d'une analyse PIONA globale (Paraffines, Isoparaffines, Oléfines, Naphtènes et Aromatiques). L'extension à d'autres types de naphtas (naphtas Straight Run, naphtas de Coker, naphtas hydrotraités, etc.) est très aisée.

Abstract* — *Reconstruction of Petroleum Feedstocks by Entropy Maximization. Application to FCC Gasolines — *In the petroleum industry, the oil fractions are usually complex mixtures containing several hundreds up to several millions of different chemical species. For this reason, even the most powerful analytical tools do not allow to separate and to identify all the species that are present. Hence, petroleum fractions are currently characterized either by using average macroscopic descriptors (density, elemental analyses, Nuclear Magnetic Resonance, etc.) or by using separative techniques (distillation, gas or liquid chromatography, mass spectrometry, etc.), which quantify only a limited number of families of molecules however. Reconstruction methods for the petroleum cuts are numerical tools, which allow to evolve towards a molecular detail and which are all based on the following principle: defining simplified but consistent mixtures of chemical compounds from partial analytical data and from expert knowledge of the process under study. Thus, the reconstruction method by entropy maximization, which is proposed in*

this article, is a recent and powerful technique which allows to determine the molar fractions of a predefined set of chemical compounds by maximizing an entropic criterion and by satisfying the analytical constraints given by the modeler. This approach allows to reduce the number of degrees of freedom from several thousands (corresponding to the molar fractions of the compounds) to several tens (corresponding to the Lagrange parameters associated with the analytical constraints) and to greatly decrease the CPU time required to perform the calculations. This approach has been successfully applied to reconstruct FCC gasolines by precisely predicting the molecular composition of this type of feedstocks from a distillation and an overall PIONA analysis (Paraffins, Isoparaffins, Olefins, Naphthenes and Aromatics). The extension to other naphthas (Straight Run naphthas, Coker naphthas, hydrotreated naphthas, etc.) is straightforward.

INTRODUCTION

In the petroleum industry, the oil fractions are usually complex mixtures containing from several hundreds to several millions of different chemical species depending on their cut points. For this reason, even the most powerful analytical tools do not allow to separate and identify all the compounds that are present. Petroleum fractions are therefore characterized analytically, either by using average macroscopic descriptors (density, elemental analyses, nuclear magnetic resonance, etc.) or by using separative techniques (distillation, gas or liquid chromatography, mass spectrometry, etc.) which quantify only a limited number of families of molecules however.

To compensate for this absence of a detailed analytical description, computational techniques have been developed to arrive at a molecular-level representation of petroleum fractions. These methods, coined molecular reconstruction methods, allow to create simplified but consistent mixtures of compounds from partial analytical data by injecting expert knowledge. They are based on different strategies that depend on the type of feedstocks to rebuild. For the heaviest petroleum fractions (vacuum gas oils, atmospheric residues, vacuum residues), the main approach consists in sampling randomly different statistical distributions of structural blocks which are the basic elements of the molecules present in the mixture after reconstruction [1-8]. For the lightest petroleum cuts (gasolines, kerosenes, gas oils), the majority of the methods is based on a predefined set of compounds which is supposed to be representative of the molecules that are actually present in the fraction under investigation. Thus, the technique consists of calculating the molar fractions of the compounds in the database in order to verify the set of analytical constraints which is associated with the feedstock to be represented [9-18].

The reconstruction method described in this article is part of the latter set of strategies. The technique is based on a predefined set of compounds whose molar fractions are calculated by maximizing an information entropy criterion subject to constraints. After a description of the theory used to develop this reconstruction method by entropy maximization, an application will be presented for gasolines withdrawn from the Fluid Catalytic Cracking process (FCC).

1 DESCRIPTION OF THE MOLECULAR RECONSTRUCTION METHOD BY ENTROPY MAXIMIZATION

1.1 Concept

The feedstock reconstruction by entropy maximization is based on the Shannon's information theory developed in 1948 [19], which is still applied in many scientific fields such as quantum chemistry or electronics. From this theory, Shannon defines an entropic criterion as follows:

$$E(\mathbf{p}) = -\sum_{i=1}^{N} p_i \cdot \ln p_i \tag{1}$$

with

$$\sum_{i=1}^{N} p_i = 1 \tag{2}$$

E represents the Shannon entropy, N is the number of possible states and p_i is the probability of the state i. To summarize, Shannon's entropy is a measure of the homogeneity of the probability distribution p_i. The higher the entropy value, the more uniform the distribution is. When the criterion is associated with constraints, maximizing the Shannon entropy is equivalent to determining the most uniform distribution which verifies these constraints.

In the case of feedstock reconstruction, the probabilities p_i have to be replaced by the molar fractions x_i of the N compounds present in the petroleum mixture. The usefulness of the entropy maximization to molecular reconstruction relies on two fundamental characteristics:

– if there are no constraints (or no petroleum analyses), it is impossible to favor one compound of the database over the others. In this case, all molar fractions x_i are equal to $1/N$. The distribution is thus uniform;

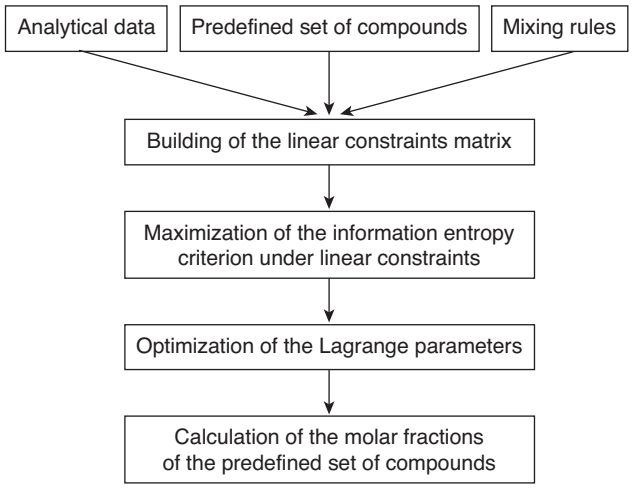

Figure 1

Scheme of feedstock reconstruction by entropy maximization.

– when constraints (or petroleum analyses) are added, the method modifies the molar fractions so as to verify these constraints but with a molar fraction distribution that stays as uniform as possible. In the special case of linear constraints, a semi-algebraic solution is possible and allows to decrease greatly the number of unknowns and the CPU time required to resolve the problem.

The different steps of the feedstock reconstruction by entropy maximization *(cf. Fig. 1)* are thus the following:

– creation of a molecular database which is as representative as possible of the set of feedstocks to rebuild (Fluid Catalytic Cracking gasolines, Straight Run naphthas, Light Cycle Oil gasoils, etc.);
– introduction of the constraints associated to the different petroleum analyses for each feedstock. When these constraints are linear or can be transformed into linear constraints, the solution space of the maximization problem can be significantly reduced and a semi-algebraic resolution of the problem can be derived;
– adjustment of the molar fractions by maximizing the information entropy criterion associated with the previous constraints.

1.2 Creation of a Predefined Database of Compounds

The creation of the initial database of compounds is extremely important because the choice of the molecules allows introducing additional information which are not necessarily present in the petroleum analyses but which are well-known to the experts. That is why each type of feedstock must have its own initial database. For example, the Light Cycle Oil gas oils (LCO) are known to contain a lot of aromatic compounds with small side chains due to the cracking reactions in the Fluid Catalytic Cracking (FCC) process. Similarly, they only have benzothiophenes and dibenzothiophenes as sulfur compounds. Introducing sulfides or aromatic compounds with long side chains in a LCO-specific database is therefore not only useless but very harmful because the future reconstructed mixtures might contain some of these compounds which do not exist in the actual LCO feedstocks. During the creation of the predefined database of compounds, it is therefore very important to verify that the compounds of the database are representative of the studied feedstocks and that the mixture obtained with a uniform distribution ($x_i = 1/N \ \forall i \in N$) has properties close to those of the feedstock. The importance of latter property will be illustrated and dealt with in more detail in the example application of the method.

To create these specific databases and avoid these problems of selection, two different approaches have been developed at IFP Energies nouvelles:

– *experimental method:* this method is used for the light petroleum cuts when detailed Gas Chromatography (GC) analyses exist and allow to determine qualitatively the different compounds that are present in the studied feedstocks. These compounds identified by the GC analyses are then used to build the database. This is the most efficient method because no selections or assumptions are needed concerning the presence or not of a compound in the initial database;
– *coupling method:* this method uses another reconstruction method, the stochastic reconstruction, to create the initial molecular database. For a more detailed description of this approach, the reader is referred to other articles [1-4; 6; 8] concerning this reconstruction technique.

An other solution consists in judiciously selecting a limited number of model compounds based on expert knowledge, as proposed by Liguras and Allen [9-11], Eckert and Vanek [16-18] or Van Geem *et al.* [20-21].

In all cases, when the initial database is finally created, it is necessary to know the various pure component properties of the selected compounds. Some properties can easily be calculated from the molecular structure of the compound, such as its chemical formula, its molecular weight, its ^{1}H Nuclear Magnetic Resonance signature (NMR), its ^{13}C NMR signature, its mass spectrometry family and its Paraffins/Isoparaffins/Olefins/Naphthenes/Aromatics (PIONA) family. For other properties, group contribution methods need to be employed. Table 1 lists the different methods used to calculate these properties. In this work, specific group contribution methods (Annex A) have been developed for the normal boiling point and the density at 20°C.

TABLE 1

Methods used to calculate molecular properties of compounds

Pure compound properties	Associated methods used for the calculation	References
Chemical formula	by inspection	
Molecular weight	by inspection	
Mass spectrometry family	by inspection	
PIONA family	by inspection	
^{1}H NMR signature	by inspection	
^{13}C NMR signature	by inspection	
Specific gravity	by group contribution method	[6], Annex A
Normal boiling point	by group contribution method	[6], Annex A

1.3 Introduction of the Petroleum Analyses

For each feedstock to be rebuilt, the introduction of its own petroleum analyses allows first of all to eliminate the compounds of the initial database that can not be present in the final mixture. To this aim, each analysis allows to filter the database in the following manner:

- *elemental analysis:* this filter eliminates all the compounds of the initial database which contain a type of atom not detected by the corresponding analysis. For example, if the sulfur content is equal to 0, all the sulfur compounds of the initial database are removed;
- *mass spectrometry:* this filter eliminates all the compounds of the initial database that belong to a mass spectrometry family that is not detected by the analysis;
- *PIONA analysis:* this filter eliminates all the compounds of the initial database that belong to a PIONA family that is not detected by the analysis;
- *^{1}H and ^{13}C NMR analyses:* this filter eliminates all the compounds of the initial database which contain a type of hydrogen or carbon that is not detected by the analyses;
- *simulated distillation:* this filter eliminates all the compounds of the initial database that have normal boiling points lower than the initial point or greater than the final point of the distillation;
- *molecular weight:* no specific filter;
- *specific gravity:* no specific filter.

When the initial database has been filtered to eliminate the "impossible" compounds, the various properties of the mixture can be obtained from the pure component properties of each compound ($P_{i,j}$) and the mixture composition (x_i) via the corresponding mixing rules (F_j) of each property j. For some analyses, such as a distillation curve or a PIONA analysis, the mixing rule is not a simple algebraic expression but a conditional expression. The calculated mixture properties can now be compared to the experimental values of each property for the petroleum fraction to be reconstructed. This therefore leads to the following set of exact constraints:

$$P_j = F_j(\mathbf{x}) \;\; \forall\, j \in J \tag{3}$$

where $\mathbf{x}$ Vector of molar fractions x_i

x_i Molar fraction of the compound i

F_j Mixing rule for property j

P_j Experimental value of property j for the petroleum fraction (constraint j)

J Total number of constraints

The total number of constraints J groups together all the constraints of the various analyses. It should indeed be stressed that the PIONA analysis for example can contain up to 5 constraints (one for each family), while the ^{13}C NMR analysis contains up to 7 constraints.

When the mixing rules are linear or can be transformed into linear mixing rules, a semi-algebraic resolution of the maximization problem can be derived *(see Sect. 1.6)*. In such a case, the mixing rules of Equation (3) can be transformed into J exact linear constraints, which are grouped together into a matrix system:

$$f_j = \sum_{i=1}^{N} f_{i,j} \cdot x_i \;\; \forall\, j \in J \tag{4}$$

where f_j Equality term for the constraint j

$f_{i,j}$ Parameter of the compound i for the constraint j

x_i Molar fraction of the compound i

N Number of compounds in the database after filtering

J Total number of constraints

In the simplest example of a linear mixing rule as Equation (4), the equality term f_j for constraint j equals the experimental value of property (P_j) for the petroleum fraction, while $f_{i,j}$ corresponds to the value of the pure component property j for compound ($P_{i,j}$). In order to account for the conditional expressions in some mixing rules (distillation curve or a PIONA analysis) and in order to improve the numerical stability, the terms f_j and $f_{i,j}$ may be transformations of these values P_j and $P_{i,j}$, respectively. Table 2 lists the set of equations used to obtain $f_{i,j}$ and f_j for all the petroleum analyses used in the feedstock reconstruction algorithm. In this table, the majority of the analyses are calculated by a simple mass balance (elemental analyses, average molecular weight, PIONA analysis, mass spectrometry and NMR analyses). However, for the simulated distillation and the specific gravity, some hypotheses must be made. For the simulated distillation, the compounds of the mixture are supposed to leave the chromatography column by increasing normal boiling points. The simulated distillation is then considered as a True Boiling Point distillation (TBP). For the specific gravity, the mixture of compounds is considered as ideal and no excess molar volume is

TABLE 2

$f_{i,j}$ and f_j for the petroleum analyses used in feedstock reconstruction

Petroleum analyses		Associated $f_{i,j}$	Associated f_j
Elemental analyses	wt%	$f_{i,j} = M_i \cdot \%X^{exp} / 100 - nX_i \cdot M_X$	$f_j = 0$
Molecular weight	–	$f_{i,j} = M^{exp} - M_i$	$f_j = 0$
Mass spectrometry	wt%	if family i = family k, $f_{i,j} = M_i \cdot (\%F_k^{exp} / 100 - 1)$ if family $i \neq$ family k, $f_{i,j} = M_i \cdot \%F_k^{exp} / 100$	$f_j = 0$
	vol%	if family i = family k, $f_{i,j} = M_i / d_i \cdot (\%F_k^{exp} / 100 - 1)$ if family $i \neq$ family k, $f_{i,j} = M_i / d_i \cdot \%F_k^{exp} / 100$	$f_j = 0$
PIONA analysis	wt%	if family i = family k, $f_{i,j} = M_i \cdot (\%F_k^{exp} / 100 - 1)$ if family $i \neq$ family k, $f_{i,j} = M_i \cdot \%F_k^{exp} / 100$	$f_j = 0$
	vol%	if family i = family k, $f_{i,j} = M_i / d_i \cdot (\%F_k^{exp} / 100 - 1)$ if family $i \neq$ family k, $f_{i,j} = M_i / d_i \cdot \%F_k^{exp} / 100$	$f_j = 0$
^{1}H NMR analysis	–	$f_{i,j} = n\mathrm{H}_{i,k} - \%\mathrm{H}_k^{exp} / 100 \cdot n\mathrm{H}_i$	$f_j = 0$
^{13}C NMR analysis	–	$f_{i,j} = n\mathrm{C}_{i,k} - \%\mathrm{C}_k^{exp} / 100 \cdot n\mathrm{C}_i$	$f_j = 0$
Distillation	wt%	if $Tb_i < Tb_k^{exp}$, $F_{i,j} = M_i \cdot (1 - \%F_k^{exp} / 100)$ if $Tb_i > Tb_k^{exp}$, $F_{i,j} = - M_i \cdot \%F_k^{exp} / 100$	$f_j = 0$
	vol%	if $Tb_i < Tb_k^{exp}$, $F_{i,j} = M_i / d_i \cdot (1 - \%F_k^{exp} / 100)$ if $Tb_i > Tb_k^{exp}$, $F_{i,j} = - M_i / d_i \cdot \%F_k^{exp} / 100$	$f_j = 0$
Specific gravity	–	$f_{i,j} = M_i \cdot (1 / d_i - 1 / d^{exp})$	$f_j = 0$

Notations

M_i	Molecular weight of the compound i (g/mol)
nX_i	Atom number of the element X in the compound i (X = C, H, S)
nA_i	Total number of atoms in the compound i
d_i	Specific gravity of the compound i (g/cm^3)
$n\mathrm{H}_{i,k}$	Number of hydrogen atoms of type k in the compound i
$n\mathrm{H}_i$	Total number of hydrogen atoms in the compound i
$n\mathrm{C}_{i,k}$	Number of carbon atoms of type k in the compound i
$n\mathrm{C}_i$	Total number of carbon atoms in the compound i
M^{exp}	Molecular weight of the mixture (g/mol)
$\%X^{exp}$	Experimental (weight or molar) fraction of the element X in the mixture (X = C, H, S)
$\%F_k^{exp}$	Experimental (weight or volume) fraction of the family k
$\%\mathrm{H}_k^{exp}$	Experimental molar fraction of the type of hydrogen k in the mixture
$\%\mathrm{C}_k^{exp}$	Experimental molar fraction of the type of carbon k in the mixture
d^{exp}	Specific gravity of the mixture (g/cm^3)
M_X	Molecular weight of the element X (X = C, H, S)
Family i	Type of the family of the compound i. It is a family MS or PIONA
Family k	Type of the currently studied family. It is a family MS or PIONA

added when the average molar volume of the mixture is calculated. Consequently, the specific gravity of the mixture is determined by the following linear mixing rule:

$$\frac{1}{d^{calc}} = \sum_{i=1}^{N} \frac{w_i}{d_i} \tag{5}$$

with d^{calc} Specific gravity of the mixture

w_i Weight fraction of the compound i

d_i Specific gravity of the compound i

N Number of compounds in the database after filtering

1.4 Entropy Maximization without Constraints

When there are no constraints (except the sum of molar fractions which must always be equal to 1), the introduction of the mole fraction balance *(Eq. 2)* by means of a Lagrange multiplier μ into the Shannon entropy criterion *(Eq. 1)* applied to the mole fractions leads to the following constrained criterion:

$$\xi(\mathbf{x}) = -\sum_{i=1}^{N} x_i \cdot \ln x_i + \mu \cdot \left(1 - \sum_{i=1}^{N} x_i\right) \tag{6}$$

At the optimum, if a solution exists, the following relations must be satisfied:

$$\frac{\partial \xi}{\partial x_i} = -1 - \ln x_i - \mu = 0 \quad \forall i \in N \tag{7}$$

Or,

$$e^{1+\mu} \cdot x_i = 1 \quad \forall i \in N \tag{8}$$

By summing Equation (8) for $i = 1$ to N, we obtain:

$$e^{1+\mu} \cdot \sum_{i=1}^{N} x_i = e^{1+\mu} = N \tag{9}$$

Introducing $\exp(1 + \mu) = N$ in Equation (8), one finally obtains:

$$x_i = \frac{1}{N} \quad \forall i \in N \tag{10}$$

The first characteristic of the entropy maximization is then verified. Without constraints, the distribution of the compounds in the database is uniform. All the molar fractions are equal to $1/N$, with N the total number of compounds of the database after filtering.

1.5 Entropy Maximization with Exact Non-linear Constraints

The introduction of exact non-linear constraints *(Eq. 3)* in the entropy criterion with the Lagrange multipliers λ_j leads to the following equation:

$$\xi(\mathbf{x}) = -\sum_{i=1}^{N} x_i \cdot \ln x_i + \mu \cdot \left(1 - \sum_{i=1}^{N} x_i\right) + \sum_{j=1}^{J} \lambda_j \cdot \left(P_j - F_j(\mathbf{x})\right) \tag{11}$$

After deriving Equation (11) with respect to all x_i, the following relations are obtained:

$$\frac{\partial \xi}{\partial x_i} = -1 - \ln x_i - \mu - \sum_{j=1}^{J} \lambda_j \cdot \frac{\partial F_j}{\partial x_i} = 0 \quad \forall i \in N \tag{12}$$

This leads to:

$$e^{1+\mu} \cdot x_i = \exp\left(-\sum_{j=1}^{J} \lambda_j \cdot \frac{\partial F_j}{\partial x_i}\right) \quad \forall i \in N \tag{13}$$

$$x_i = \frac{\exp\left(-\sum_{j=1}^{J} \lambda_j \cdot \frac{\partial F_j}{\partial x_i}\right)}{Z} \quad \forall i \in N \tag{14}$$

with

$$Z = \sum_{i=1}^{N} \exp\left(-\sum_{j=1}^{J} \lambda_j \cdot \frac{\partial F_j}{\partial x_i}\right) \tag{15}$$

Introducing Equation (14) into Equation (11) allows obtaining the following relation:

$$\xi(\mathbf{x}, \boldsymbol{\lambda}) = \ln Z + \sum_{j=1}^{J} \lambda_j \cdot P_j - \sum_{j=1}^{J} \lambda_j \cdot \left(F_j(\mathbf{x}) - \frac{1}{Z} \cdot \sum_{i=1}^{N} \left(\frac{\partial F_j}{\partial x_i} \cdot \exp\left(-\sum_{j=1}^{J} \lambda_j \cdot \frac{\partial F_j}{\partial x_i}\right)\right)\right) \tag{16}$$

The entropy criterion of Equation (11) with J non-linear constraints contained $N + J$ unknowns, *i.e.* the Lagrange multipliers λ_j of each constraint and the mole fraction x_i of each compound. As the constraints are non-linear, the reformulated problem in Equation (16) may still contain up to $N + J$ unknowns, depending on the non-linearities in the mixing rules F_j. Maximization of the entropy criterion under non-linear constraints is therefore a non-trivial task due to the high dimension of the solution space.

1.6 Entropy Maximization with Exact Linear Constraints

The introduction of exact linear constraints in the entropy criterion with the Lagrange multipliers λ_j allows obtaining the following equation:

$$\xi(\mathbf{x}) = -\sum_{i=1}^{N} x_i \cdot \ln x_i + \mu \cdot \left(1 - \sum_{i=1}^{N} x_i\right) + \sum_{j=1}^{J} \lambda_j \cdot \left(f_j - \sum_{i=1}^{N} x_i \cdot f_{i,j}\right) \tag{17}$$

After deriving Equation (17) with respect to all x_i, the following relations are found:

$$\frac{\partial \xi}{\partial x_i} = -1 - \ln x_i - \mu - \sum_{j=1}^{J} \lambda_j \cdot f_{i,j} = 0 \quad \forall i \in N \tag{18}$$

This leads to:

$$e^{1+\mu} \cdot x_i = \exp\left(-\sum_{j=1}^{J} \lambda_j \cdot f_{i,j}\right) \quad \forall i \in N \tag{19}$$

$$x_i = \frac{\exp\left(-\sum_{j=1}^{J} \lambda_j \cdot f_{i,j}\right)}{Z} \quad \forall i \in N \tag{20}$$

with

$$Z = \sum_{i=1}^{N} \exp\left(-\sum_{j=1}^{J} \lambda_j \cdot f_{i,j}\right) \tag{21}$$

Introducing Equation (20) into Equation (17) allows finally to obtain the following relation:

$$\xi(\boldsymbol{\lambda}) = \ln Z + \sum_{j=1}^{J} \lambda_j \cdot f_j \tag{22}$$

Maximizing the entropy criterion under J exact linear constraints (*Eq. 17* with $N + J$ unknowns) is now elegantly reduced to maximizing Equation (22) which only contains J unknowns, the Lagrange multipliers λ_j of each constraint. Hence, the solution space of the maximization problem has been significantly reduced from several thousands (the molar fractions of the compounds contained in the initial database) to a few tens (the Lagrange parameters associated to the constraints). The maximization of the latter non-linear equation can be done by classic optimization methods such as, for example, the conjugate gradient method. When the values of λ_j are determined at the optimum, the molar fractions x_i of the compounds can be calculated using Equations (20) and (21).

To illustrate the entropy maximization with exact linear constraints, an example is given using a database containing ten compounds. The molecular weights of these compounds are respectively equal to 100, 110, 120, 130, 140, 150, 160, 170, 180 and 190 g/mol. The entropy maximization is then used to rebuild three mixtures which have respectively an average molecular weight of 115 g/mol (example 1), 142 g/mol (example 2) and 178 g/mol (example 3). After optimization, the obtained values of λ, the Lagrange multiplier associated with the constraint on the average molecular weight, are respectively 0.049195 (example 1), 0.003644 (example 2) and –0.059655 (example 3). Using Equations (20) and (21) finally allows to obtain the molar fractions of each compound for the three examples *(Tab. 3 and Fig. 2)*.

If the molar distribution of the database would have been uniform ($x_i = 0.1$, $1 \leq i \leq 10$), the obtained mixture would have had an average molecular weight of 145 g/mol, which will be referred to as the equimolar molecular weight of the database. As can be observed, two cases appear depending on the average molecular weight to be fitted to:

– if the average molecular weight to be fitted to is very different from the equimolar molecular weight (examples 1 and 3), the absolute value of λ is relatively high and the entropy maximization largely favors some compounds that are extreme with respect to their properties. This is due to the exponential term of Equation (20);

TABLE 3

Results of the 3 examples of reconstruction with exact constraints

		Molar fractions		
Compounds	Molecular weight (g/mol)	Example 1	Example 2	Example 3
1	100	0.3914	0.1172	0.0021
2	110	0.2393	0.1130	0.0038
3	120	0.1463	0.1089	0.0069
4	130	0.0895	0.1050	0.0126
5	140	0.0547	0.1013	0.0228
6	150	0.0334	0.0977	0.0414
7	160	0.0205	0.0942	0.0752
8	170	0.0125	0.0908	0.1366
9	180	0.0076	0.0875	0.2481
10	190	0.0047	0.0844	0.4504
Optimized parameter λ		0.049195	0.003644	–0.059655
Average molecular weight		115 g/mol	142 g/mol	178 g/mol

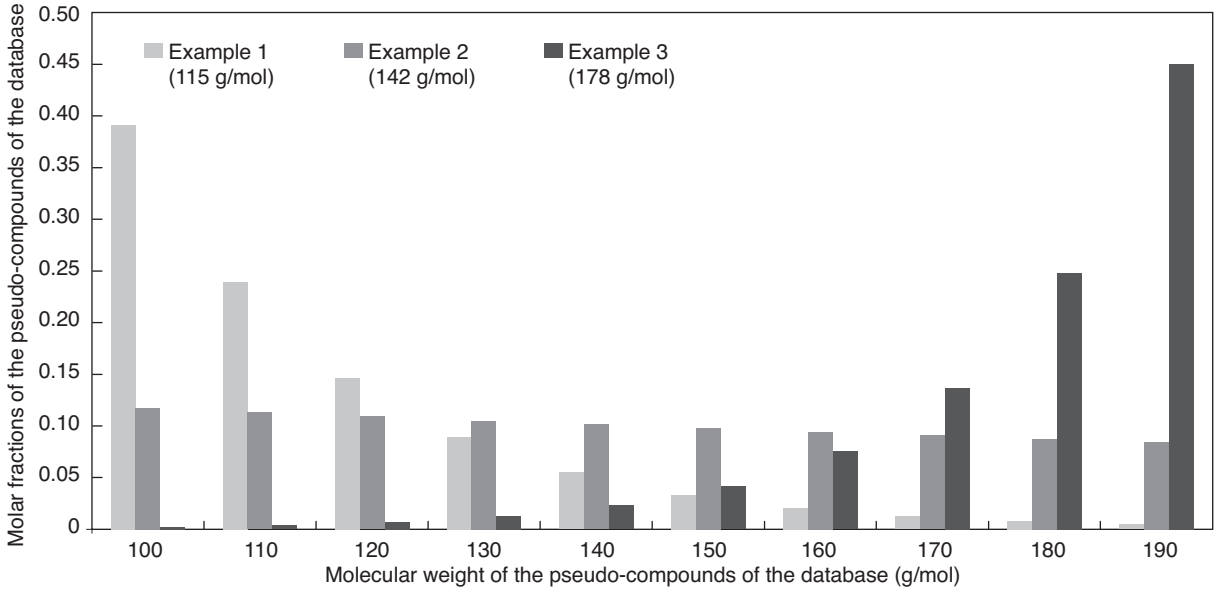

Figure 2

Results of the 3 examples of reconstruction with exact constraints.

– if the average molecular weight to be fitted to is close to the equimolar molecular weight (example 2), the value λ is near zero and the mole fraction distribution is almost uniform.

It is for this reason that it is important to stress that the initial database of compounds must be as characteristic as possible of the feedstocks to be rebuilt. The database should also have mixture properties (based on an equimolar mixture) as close as possible of the properties of the feedstocks to be reconstructed in order to avoid the appearance of a limited number of dominating molecules.

1.7 Entropy Maximization with Linear Constraints Containing Uncertainties

Up to now, all the constraints have been qualified as "exact" because the entropy maximization method forced the system to exactly fit the different constraints. But analytical data also contain measurement errors. Moreover, the hypotheses used to determine some of the properties of the rebuilt mixture also introduce some deviations. Trying to perfectly fit to the experimental data is therefore completely pointless. This is why Equation (17) can be modified to handle K constraints with uncertainties in the following manner:

$$\xi(\mathbf{x}) = -\sum_{i=1}^{N} x_i \cdot \ln x_i + \mu \cdot \left(1 - \sum_{i=1}^{N} x_i\right) + \sum_{k=1}^{K} -\frac{1}{2} \cdot \frac{\left(f_k - \sum_{i=1}^{N} x_i \cdot f_{i,k}\right)^2}{\sigma_k^2} \quad (23)$$

where σ_k Uncertainty on the constraint k

K Total number of constraints with uncertainties

The system resolution is then similar to the case with exact linear constraints. The intermediate equations are:

$$\frac{\partial \xi}{\partial x_i} = -1 - \ln x_i - \mu + \sum_{k=1}^{K} \frac{f_{i,k}}{\sigma_k} \cdot \frac{f_k - \sum_{i=1}^{N} x_i \cdot f_{i,k}}{\sigma_k} = 0 \ \forall\ i \in N \quad (24)$$

If we define the normalized residuals as:

$$\varepsilon_k = \frac{f_k - \sum_{i=1}^{N} x_i \cdot f_{i,k}}{\sigma_k} \quad \forall k \in K \quad (25)$$

we have finally:

$$e^{1+\mu} \cdot x_i = \exp\left(\sum_{k=1}^{K} \varepsilon_k \cdot \frac{f_{i,k}}{\sigma_k}\right) \quad \forall i \in N \quad (26)$$

$$x_i = \exp\left(\sum_{k=1}^{K} \varepsilon_k \cdot \frac{f_{i,k}}{\sigma_k}\right) / Z \quad \forall i \in N \quad (27)$$

$$Z = \sum_{i=1}^{N} \exp\left(\sum_{k=1}^{K} \varepsilon_k \cdot \frac{f_{i,k}}{\sigma_k}\right) \quad (28)$$

The new non-linear equation to maximize is then:

$$\xi(\boldsymbol{\varepsilon}) = \ln Z + \sum_{k=1}^{K} \left(\frac{1}{2} \cdot \varepsilon_k^2 - \frac{f_k}{\sigma_k} \cdot \varepsilon_k\right) \quad (29)$$

Maximizing Equation (29) allows to determine the values of the normalized residuals ε_k at the optimum. The molar fractions x_i can then be calculated by using Equations (27) and (28).

Now, taking again the example of entropy maximization with the database containing ten compounds, it is possible to introduce uncertainties on the molecular weight. Recall that the initial database is composed of 10 compounds which have respectively a molecular weight of 100, 110, 120, 130, 140, 150, 160, 170, 180 and 190 g/mol. If the mixture is equimolar, its average molecular weight is equal to 145 g/mol (termed the equimolar molecular weight of the database). The entropy maximization is then used to rebuild three mixtures which have an average molecular weight of 160 g/mol but with different values of uncertainties defined by σ: 1 (example 4), 10 (example 5) and 100 (example 6). After optimization, the obtained values of ε are respectively 0.0192646 (example 4), 0.16894221 (example 5) and 0.13857154 (example 6). Using Equations (27) and (28) finally allows to obtain the molar fractions of the compounds for each of three examples *(Tab. 4 and Fig. 3)*.

As can be observed, the smaller the value of σ, the more the system is constrained and tends to the value of average molecular weight to be fitted to (160 g/mol). Conversely, the higher the value of σ, the less the system is constrained and

TABLE 4

Results of the 3 examples of reconstruction with constraints with uncertainties

		Molar fractions		
Compounds	Molecular weight (g/mol)	Example 4	Example 5	Example 6
1	100	3.622E-02	4.168E-02	9.388E-02
2	110	4.392E-02	4.935E-02	9.519E-02
3	120	5.325E-02	5.843E-02	9.652E-02
4	130	6.456E-02	6.918E-02	9.787E-02
5	140	7.828E-02	8.191E-02	9.923E-02
6	150	9.491E-02	9.699E-02	1.006E-01
7	160	1.151E-01	1.148E-01	1.020E-01
8	170	1.395E-01	1.360E-01	1.034E-01
9	180	1.692E-01	1.610E-01	1.049E-01
10	190	2.051E-01	1.906E-01	1.063E-01
Parameter σ		1	10	100
Optimized parameter ε		0.0192646	0.16894221	0.13857154
Average molecular weight		160 g/mol	158 g/mol	146 g/mol

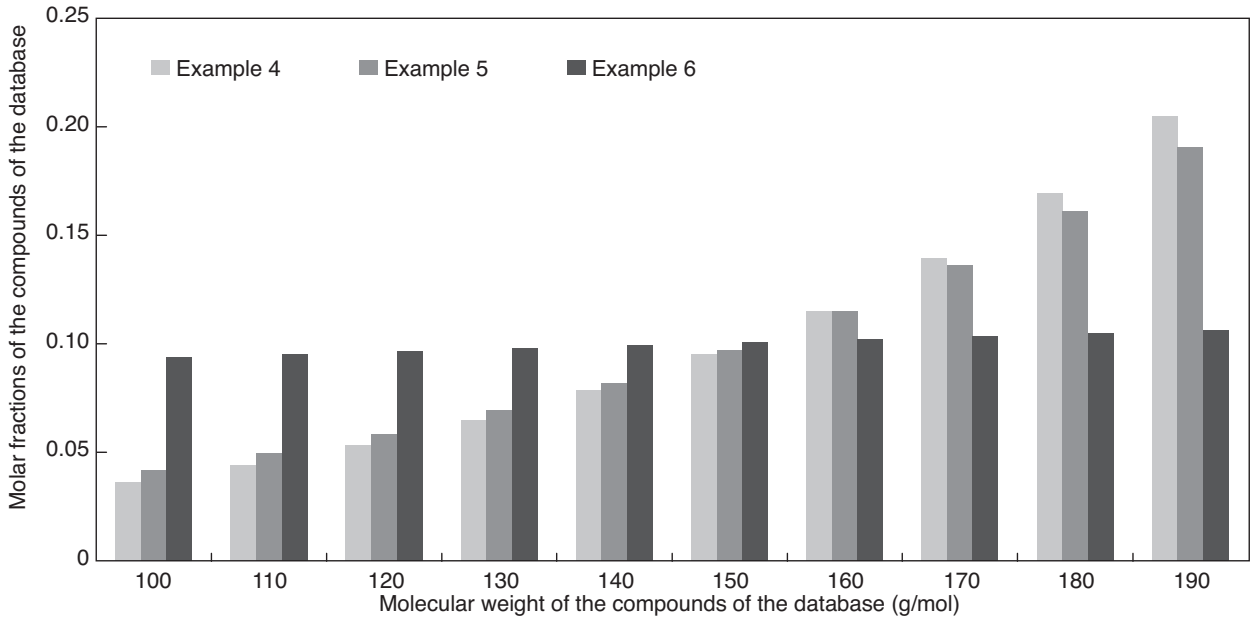

Figure 3

Results of the 3 examples of reconstruction with constraints with uncertainties.

the more it tends to the equimolar mixture (which is the base case when there are no constraints).

1.8 Entropy Maximization with Introduction of Normal Distributions

In the cases of exact linear constraints or linear constraints with uncertainties, the entropy maximization method tends to favor the compounds with extreme properties to the detriment of more average components. This is due to the exponential term of the functions that allow to calculate the molar fractions *(cf. Eq. 20 and 27)*. However, it is known that the most of distributions of compounds in petroleum fractions are not exponential but rather of a Gaussian type with a majority of components with average properties and few extreme compounds. To force some of the characteristics to follow a specific statistical distribution, it is possible to add a new constraint for each of these characteristics. In case one wants to impose a normal distribution on a given characteristic, the following constraint has to be included in addition to its equality constraint (exact or with uncertainties):

$$\left(\sigma^\circ{}_m\right)^2 = \sum_{i=1}^{N} x_i \cdot \left(f_{i,m} - f_m\right)^2 \tag{30}$$

with $\sigma^\circ{}_m$ Standard deviation of the normal distribution linked to constraint m.

The general equation to maximize is then the following:

$$\begin{aligned}\xi(\mathbf{x}) = &-\sum_{i=1}^{N} x_i \cdot \ln x_i + \mu\left(1 - \sum_{i=1}^{N} x_i\right) \\ &+ \sum_{j=1}^{J} \lambda_j \cdot \left(f_j - \sum_{i=1}^{N} x_i \cdot f_{i,j}\right) \\ &+ \sum_{k=J+1}^{J+K} -\frac{1}{2} \cdot \frac{\left(f_k - \sum_{i=1}^{N} x_i \cdot f_{i,k}\right)^2}{\sigma_k^2} \\ &+ \sum_{m \in [L]} \nu_m \cdot \left(\left(\sigma^\circ{}_m\right)^2 - \sum_{i=1}^{N} x_i \cdot \left(f_m - f_{i,m}\right)^2\right)\end{aligned} \tag{31}$$

with ν_m The Lagragian parameter corresponding to the normal distribution constraint of characteristic m

$[L]$ A subspace of the J exact constraints and the K constraints with uncertainties

After intermediate calculations, the entropy maximization amounts to maximizing the non-linear equation:

$$\xi\left(\boldsymbol{\lambda}, \boldsymbol{\varepsilon}, \boldsymbol{\nu}\right) = \ln Z + \sum_{j=1}^{J} \lambda_j \cdot f_j + \sum_{k=J+1}^{J+K} \left(\frac{1}{2} \cdot \varepsilon_k^2 - \frac{f_k}{\sigma_k} \cdot \varepsilon_k\right) + \sum_{m \in [L]} \nu_m \cdot \left(\sigma^\circ{}_m\right)^2 \tag{32}$$

with:

$$Z = \sum_{i=1}^{N} \exp\left(-\sum_{j=1}^{J} \lambda_j \cdot f_{i,j} + \sum_{k=J+1}^{J+K} \varepsilon_k \cdot \frac{f_{i,k}}{\sigma_k} - \sum_{m\in[L]} \nu_m \cdot \left(f_{i,m} - f_m\right)^2 \right) \quad (33)$$

Maximization of the function ξ is performed by a classic numerical method such as the conjugate gradient method. After optimization, the resulting parameters ($\boldsymbol{\lambda}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\nu}$) allow to calculate the molar fractions by means of the following equation:

$$= \exp\left(-\sum_{j=1}^{J} \lambda_j \cdot f_{i,j} + \sum_{k=J+1}^{J+K} \varepsilon_k \cdot \frac{f_{i,k}}{\sigma_k} - \sum_{m\in[L]} \nu_m \cdot \left(f_{i,m} - f_m\right)^2 \right) / Z \quad (34)$$

$\forall i \in N$

To illustrate this new type of constraints, the example of the molecular weight can be taken again. In the present case, the target average molecular weight of the mixture is exactly 142 g/mol. A new constraint on the molecular weight is also added to force the system to have a normal distribution. The entropy maximization is then used to rebuild three mixtures with different standard deviations on the molecular weight: $\sigma° = 5$ (example 7), $\sigma° = 10$ (example 8) and $\sigma° = 20$ (example 9). The obtained molar fractions and the optimized parameters λ and ν are given in Table 5 and Figure 4.

As can be observed, the addition of this new constraint allows to force the system to have a normal distribution *(Fig. 4)* instead of the non-realistic exponential distribution *(Fig. 2)*. However, this modification increases the number of parameters to determine. Thus, the parameter λ of the examples 1 to 3 is replaced by the couple (λ, ν) of the examples 7 to 9.

TABLE 5

Results of the 3 examples of reconstruction with normal repartitions

		Molar fractions		
Compounds	Molecular weight (g/mol)	Example 7	Example 8	Example 9
1	100	0.0000	0.0001	0.0287
2	110	0.0000	0.0024	0.0640
3	120	0.0001	0.0355	0.1144
4	130	0.0432	0.1942	0.1644
5	140	0.7151	0.3910	0.1898
6	150	0.2400	0.2897	0.1761
7	160	0.0016	0.0790	0.1313
8	170	0.0000	0.0079	0.0787
9	180	0.0000	0.0003	0.0379
10	190	0.0000	0.0000	0.0147
Parameter $\sigma°$		5	10	20
Optimized parameter λ		–0.007783	0.000000	0.000914
Optimized parameter ν		0.019490	0.005000	0.001093
Average molecular weight		142 g/mol	142 g/mol	142 g/mol

It is interesting to notice that the introduction of normal distributions by inclusion of additional constraints is only required when no information is given concerning the distribution of the compounds by volatility (normal boiling point or carbon number for instance). If a distillation curve is introduced as input data, this type of constraints is not necessary because the distillation already contains, by definition, some information on the distribution by volatility.

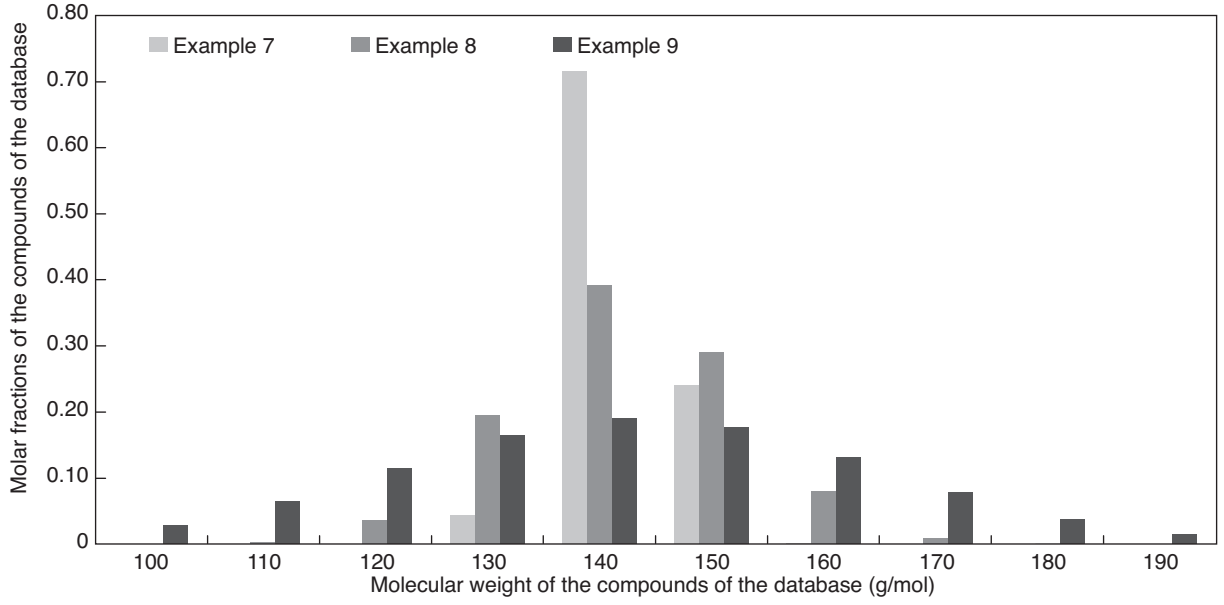

Figure 4

Results of the 3 examples of reconstruction with normal repartitions.

1.9 Preliminary Conclusions

The entropy maximization algorithm is an elegant method which allows to rebuild different types of feedstocks from partial analytical information. For that, a Shannon's entropy criterion is used, which is subject to the analytical constraints. Maximized under linear constraints (exact or with uncertainties), the mathematical expression of this criterion can be reorganized to allow for a semi-algebraic resolution of the problem with a decrease of the number of unknowns from several thousands (the molar fractions of the compounds contained in the initial database) to a few tens (the Lagrange parameters associated to the constraints). The decrease of the number of unknowns limits the CPU time required to reach the solution. The various theoretical aspects of this method have been presented above:

– without information (or without constraints), the obtained solution is an equimolar distribution of the set of the predefined compounds;
– when some information is introduced, the distribution is distorted in order to satisfy the new constraints. The information can be injected under the form of exact linear constraints *(see Sect. 1.6)* or constraints containing uncertainties *(see Sect. 1.7)*.

This methodology has however one main drawback that needs to be recalled here. Indeed, because the calculation of the molar fractions requires the use of exponentials, it is possible to obtain some unrealistic distributions leading the presence of a few predominating compounds, all the others having molar fractions close to 0. To avoid this problem, one solution consists in having an initial set of compounds whose properties (when the distribution is considered as equimolar) are close to the properties to be fitted. Another possibility proposed in this article is to use additional constraints that force the system to have normal distributions instead of exponential distributions *(see Sect. 1.8)*.

In order to illustrate this methodology for the feedstock reconstruction, different applications can be proposed such as LCO reconstruction for example [6-7]. In this paper, the reconstruction of various FCC gasolines by entropy maximization is presented. The advantage of FCC gasolines is that this type of petroleum feedstock can be analytically characterized by Gas Chromatography. The latter method allows to directly define the compounds that need to be present in the initial database without using some expert hypotheses. Moreover, the validation of the concept of entropy maximization is more practical in this application because there is a direct analytical access to the molecular composition of the various FCC feedstocks, allowing for a detailed comparison.

2 APPLICATION TO THE RECONSTRUCTION OF FCC GASOLINES

The hydrodesulfurization of the FCC gasolines has become an important field of research as the sulfur content has been decreased drastically in the on-road gasolines: 350 wt ppm before 2000, 50 wt ppm from 2005 and 10 wt ppm in 2009 for the European Union countries for example. Indeed, the largest part (over 90 wt%) of sulfur in gasoline pool comes from FCC gasoline. Refiners have then to hydrodesulfurize this fraction selectively to remove sulfur while minimizing olefins hydrogenation in order to maintain a high octane number in the final product [22].

In order to simulate the hydrodesulfurization process, it is necessary to correctly describe the FCC gasolines on a molecular level. For that, it is possible to use some Gas Chromatography techniques that allow to detect and to quantify the different chemical compounds present in this type of petroleum fractions. However, in some cases, the available information on the gasoline can be less detailed and the only accessible data concerns partial analyses such as distillation or PIONA.

In this particular configuration, the entropy maximization can be applied to rebuild FCC gasolines and to determine their molar composition. In this article, we propose to show more precisely how the reconstruction method can be applied to rebuild FCC gasolines using only the simulated distillation and the overall PIONA information. For that, it is necessary to correctly define the database of compounds that will be used during the reconstruction.

2.1 Creation of an Initial Database of Representative Compounds by Means of a Qualitative Approach (Database G)

In order to rebuild the initial database of molecules, a first possibility consists in determining analytically which are the characteristic molecules present in the FCC gasolines. It is therefore necessary to know the molecular composition of some FCC gasolines so as to identify the various chemical compounds which can be encountered in this type of oil fractions. This first operation can be carried out using a Gas Chromatography analysis. Table 6 provides general information on 15 FCC gasolines used for the definition of the initial database of molecules.

From these FCC gasolines, it is possible to build an average gasoline by calculating an equimolar mixture of the 15 gasolines defined in Table 6. This mixture allows to cover a large domain of compounds and to include the smallest and the largest molecules that can be found in FCC gasoline cuts. The compounds, detected by Gas Chromatography and having a non-zero molar fraction are then added to the initial database of molecules. Once these operations carried out, the database, referred to as database G, contained 230 different compounds. Table 7 lists the names of the first 20 compounds and the last compound of this database G. In the absence of information, *i.e.* in the absence of simulated distillation and PIONA, the entropy maximization method

TABLE 6

General information on the 15 FCC gasolines used for the definition of the initial database of molecules

	Simulated distillation (°C)	Density at 15°C (g/cm^3)	RON indices (-)	MON indices (-)	Hydrogen content (wt%)	Average *Mw* (g/mol)	P/I/O/N/A (wt%)
#1	–6.0-244.0	0.7390	91.8	79.8	13.53	99.4	4.3/24.4/35.7/6.0/29.7
#2	4.0-243.0	0.7323	91.6	80.0	13.66	96.0	4.3/28.8/32.4/5.7/28.8
#3	4.0-241.0	0.7318	91.7	81.2	13.66	95.6	4.3/28.8/32.3/5.8/28.8
#4	0.0-181.0	0.7129	92.0	79.5	14.16	90.2	4.3/28.1/47.6/6.5/13.5
#5	24.0-202.0	0.7327	94.0	80.8	13.62	94.0	3.6/22.9/44.0/5.7/23.8
#6	15.0-201.0	0.7297	93.6	80.8	13.67	92.9	3.7/23.3/44.3/5.8/22.9
#7	0.0-204.0	0.7138	92.8	80.5	14.15	89.7	4.6/32.5/37.6/7.6/17.6
#8	0.0-201.0	0.7181	92.2	80.5	14.11	90.8	5.8/32.1/35.4/8.2/18.5
#9	–8.0-203.0	0.7189	92.0	79.6	14.08	92.5	4.6/25.8/47.2/6.9/15.5
#10	23.0-190.0	0.7354	91.8	79.3	13.84	98.0	4.5/24.2/44.1/8.4/18.8
#11	54.0-189.0	0.7470	90.0	78.7	13.70	102.8	4.2/24.2/41.0/9.6/21.00
#12	2.0-178.0	0.7102	93.3	80.9	14.17	87.9	4.3/29.8/43.1/7.7/15.0
#13	0.0 -145.0	0.6769	94.0	82.0	14.91	79.3	5.4/35.6/48.9/5.8/4.3
#14	2.0-239.0	0.7510	92.8	80.9	13.31	99.9	4.0/23.8/34.6/6.4/31.3
#15	–4.0-252.0	0.7462	92.7	81.3	13.36	97.7	3.9/25.8/32.0/7.1/31.3

TABLE 7

List of the compounds contained in the database for FCC gasolines reconstruction

Database G (qualitative)		Database G+ (quantitative)	
Index	Name of the compound	Index	Name of the compound
#1	1,1,2-trimethylcyclopentane	#1	1,1,2-trimethylcyclopentane
#2	1,1,3-trimethylcyclopentane	#2	1,1,2-trimethylcyclopentane
#3	1,1-dimethylcyclohexane	#3	1,1,2-trimethylcyclopentane
#4	1,1-dimethylcyclopentane	#4	1,1,2-trimethylcyclopentane
#5	1,2,3,4-tetramethylbenzene	#5	1,1,2-trimethylcyclopentane
#6	1,2,3,5-tetramethylbenzene	#6	1,1,2-trimethylcyclopentane
#7	1,2,3-trimethylbenzene	#7	1,1,2-trimethylcyclopentane
#8	1,2,4,5-tetramethylbenzene	#8	1,1,2-trimethylcyclopentane
#9	1,2,4-trimethylbenzene	#9	1,1,2-trimethylcyclopentane
#10	1,2-butadiene	#10	1,1,2-trimethylcyclopentane
#11	1,2-dimethyl-3-ethylbenzene	#11	1,1,2-trimethylcyclopentane
#12	1,2-dimethyl-4-ethylbenzene	#12	1,1,2-trimethylcyclopentane
#13	1,3,5-trimethylbenzene	#13	1,1,2-trimethylcyclopentane
#14	1,3-butadiene	#14	1,1-dimethylcyclohexane
#15	1,3-diethylbenzene	#15	1,1-dimethylcyclohexane
#16	1,3-dimethyl-4-ethylbenzene	#16	1,1-dimethylcyclohexane
#17	1,3-dimethyl-5-ethylbenzene	#17	1,1-dimethylcyclohexane
#18	1,4-dimethyl-2-ethylbenzene	#18	1,1-dimethylcyclohexane
#19	1,5-dimethylcyclopentene	#19	1,1-dimethylcyclohexane
#20	1,C2,T4-trimethylcyclopentane	#20	1,1-dimethylcyclohexane
…	…	…	…
#230	trimethylcyclopentene	#50 003	trimethylcyclopentene

supposes an equimolar mixture. Each of the 230 compounds of the database then has a molar fraction equal to 1/230 or 4.348×10^{-3}. The corresponding mixture thus has molar fractions that are completely different when compared to a standard FCC gasoline. Hence, the properties of this mixture are far from those of typical FCC gasolines, as shown in Table 8. Indeed, the average density and the average molecular weight are too high in comparison to a conventional FCC gasoline, while both the distillation data and the carbon number distribution are very atypical. This point is particularly important because, even though the initial database of molecules contains all compounds which may actually be present in FCC gasolines, its mixture properties in the absence of information are very different from those of typical FCC gasolines. When using this database G to reconstruct various FCC gasolines by means of

the above-presented entropy maximization algorithm with constraints containing uncertainties, the introduction of simulated distillation and overall PIONA analyses was not sufficient to counter this initial bias. After reconstruction, the resulting mixture typically contained a limited number of predominant molecules, the others having a very low molar fraction. Hence, although the resulting mixtures have the same properties as the FCC gasoline to be reconstructed, their molecular composition is very different from the experimentally measured composition *(Fig. 5)*. The use

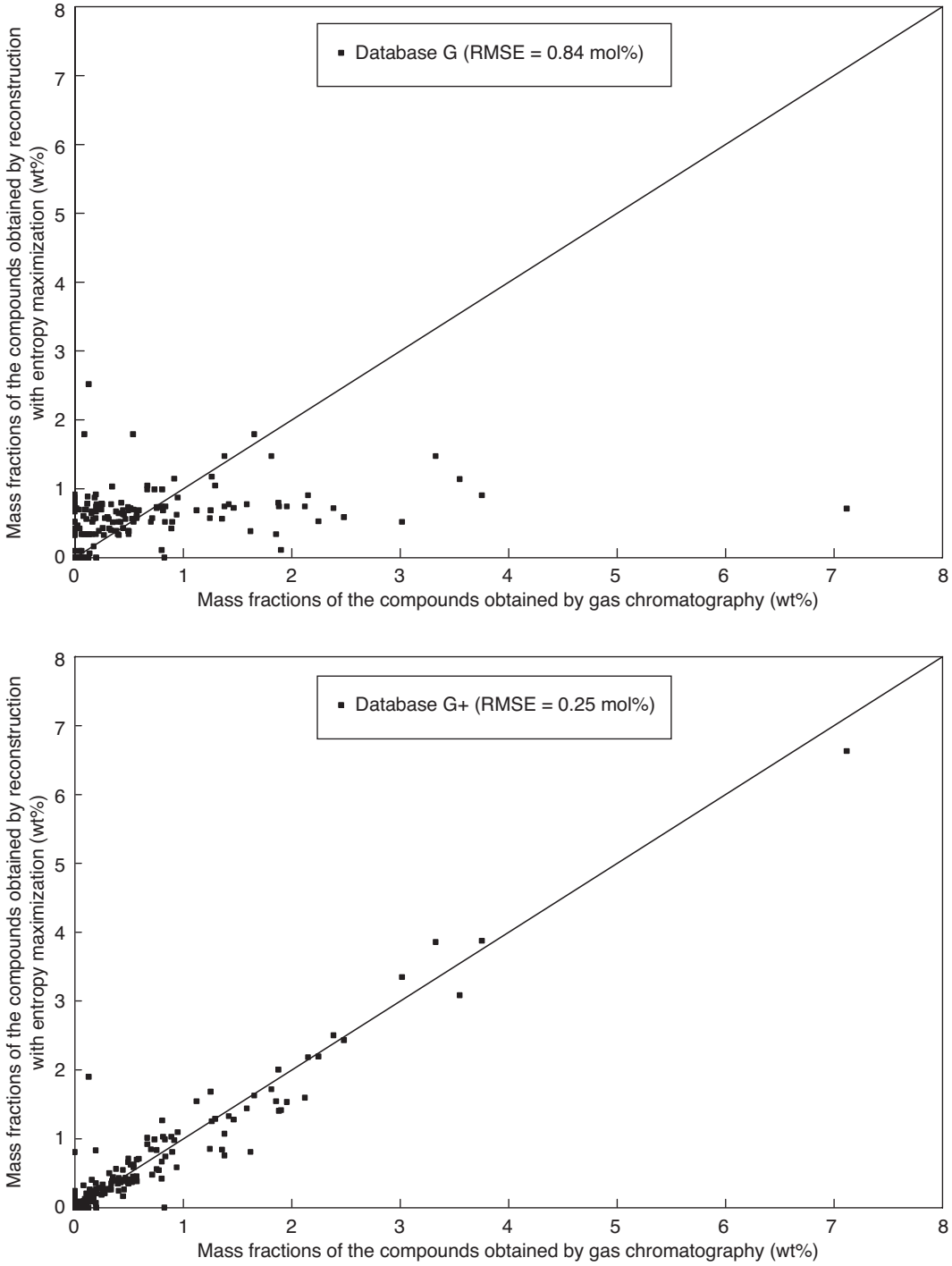

Figure 5

Comparison of the prediction for the gasoline #A depending on the choice of the initial database G or G+.

TABLE 8

Properties for the equimolar distribution of the databases G and G+

		Properties of the database G when equimolar distribution	Properties of the database G+ when equimolar distribution
Density at 15°C	(g/cm^3)	0.7701	0.7326
Average molecular weight	(g/mol)	117.5	94.3
Elemental analysis			
Carbon content	(wt%)	86.24	86.38
Hydrogen content	(wt%)	13.40	13.60
Sulfur content	(wt%)	0.36	0.02
PIONA			
Paraffins content	(wt%)	6.61	4.35
Isoparaffins content	(wt%)	27.77	26.37
Olefins content	(wt%)	28.87	39.03
Naphthenes content	(wt%)	11.65	6.90
Aromatics content	(wt%)	25.10	23.34
Distillation			
0 wt%	(°C)	–31.7	–31.7
5 wt%	(°C)	47.0	23.1
10 wt%	(°C)	69.2	33.9
20 wt%	(°C)	92.1	49.3
30 wt%	(°C)	111.1	68.8
40 wt%	(°C)	125.0	80.8
50 wt%	(°C)	144.3	98.5
60 wt%	(°C)	174.0	112.0
70 wt%	(°C)	189.1	134.6
80 wt%	(°C)	209.9	147.2
90 wt%	(°C)	236.7	175.8
95 wt%	(°C)	254.3	200.4
100 wt%	(°C)	267.6	267.6
Carbon distribution			
C_3-C_5	(wt%)	5.21	20.55
C_6-C_8	(wt%)	40.75	55.18
C_9-C_{11}	(wt%)	34.57	22.49
C_{12}-C_{14}	(wt%)	16.95	1.78
C_{15}-C_{17}	(wt%)	1.58	0.00
C_{18}+	(wt%)	0.94	0.00

of a simple qualitative database, such as this database G, is therefore not advised for molecular reconstruction by entropy maximization, unless normal distributions are judiciously introduced on some or all of the properties. In the latter case, the user also has to correctly choose the value for the standard deviation of the normal distribution linked to each constraint. This choice being tricky, an alternative database has been developed based on a quantitative approach.

2.2 Creation of an Initial Database of Representative Compounds by Means of a Quantitative Approach (Database G+)

To build a quantitative initial database of molecules, the same compound is introduced several times in the database. The frequency of each compound is directly given by the product of its molar fraction in the average gasoline and the selected total number of molecules in the quantitative initial database. For example, the total number of molecules in the quantitative initial database can be fixed at 50 000. The molar fraction of each of the 230 compounds is thus multiplied by 50 000 to obtain its frequency F and each compound is introduced F times to finally obtain a database of 50 003 molecules, referred to as database G+, as shown in Table 7. The fact that the database contains 50 003 molecules instead of 50 000 is due to round-offs. The distribution in this database of 50 003 molecules is uniform with respect to these 50 003 molecules, but with respect to the 230 compounds of the database the actual molar distribution of the average FCC gasoline is obtained. Indeed, when a mixture is rebuilt in absence of information (*i.e.* each of 50 003 molecules of the database G+ then has a molar fraction of 1/50003 or 1.9999×10^{-5}), its characteristics are this time equal to the properties of the average gasoline *(Tab. 8)*, which are, by construction, very close to the properties of typical FCC gasolines. As will be detailed in the next section, when starting from this new database G+ to reconstruct various FCC gasolines by means of the above-presented entropy maximization algorithm with constraints containing uncertainties, the mere use of simulated distillation and overall PIONA analyses allows to obtain a very good molecular representation of the actual feedstock. This is totally different from the results obtained with database G as is illustrated in Figure 5.

2.3 Prediction of the Molecular Composition of a FCC Gasoline from its Distillation and PIONA

From the quantitative initial database G+ of FCC gasoline compounds, it is possible to predict the molecular composition of FCC gasolines by simply using its distillation data and its overall PIONA analysis. For that, the reconstruction by entropy maximization is employed using linear constraints with uncertainties as described in the theoretical part of this article. Three different gasolines (1 present in the equimolar mixture used to build the initial database, together with 2 new gasolines) have been rebuilt according to this algorithm. Table 9 gives the input data (distillation and overall PIONA analysis) for these 3 gasolines.

The FCC gasolines have initially been rebuilt by using both databases G and G+. After reconstruction, the obtained mixtures have distillation and PIONA analyses that are very close to the analytical data of the gasoline,

TABLE 9

Analytical data used to rebuild 3 different FCC gasolines

	Gasoline #A	Gasoline #B	Gasoline #C (1)
Simulated distillation			
0 wt%	54°C	54°C	0°C
5 wt%	55°C	69°C	23°C
10 wt%	69°C	73°C	23°C
20 wt%	75°C	93°C	30°C
30 wt%	89°C	112°C	37°C
40 wt%	99°C	128°C	40°C
50 wt%	112°C	144°C	55°C
60 wt%	123°C	164°C	61°C
70 wt%	137°C	177°C	70°C
80 wt%	145°C	194°C	81°C
90 wt%	162°C	213°C	99°C
95 wt%	169°C	230°C	112°C
100 wt%	189°C	254°C	145°C
PIONA			
Paraffins content	4.21 wt%	3.75 wt%	5.38 wt%
Isoparaffins content	24.25 wt%	20.11 wt%	35.65 wt%
Olefins content	41.02 wt%	32.11 wt%	48.93 wt%
Naphthenes content	9.57 wt%	7.04 wt%	5.76 wt%
Aromatics content	20.95 wt%	36.99 wt%	4.28 wt%

(1) Gasoline #C corresponds to Gasoline #13 used to build the initial database of compounds

TABLE 10

Statistical evaluation of the results obtained by entropy maximization of various FCC gasolines

Gasoline index	#A	#A	#B	#C
Used database	database G	database G+	database G+	database G+
RSS ((mol%)^2)	124.5	11.0	19.8	24.7
RMSE (mol%)	0.84	0.25	0.31	0.41
MPE (-)	1.10	0.03	–0.05	0.01
AAD (mol%)	0.49	0.15	0.13	0.18

RSS: Residual Sum of Squares ($RSS = \sum (x_i - \hat{x}_i)^2$)

RMSE: Root Mean Squared Error ($RMSE = \sqrt{\frac{\sum (x_i - \hat{x}_i)^2}{N}}$)

MPE: Mean Percentage Error ($MPE = \frac{1}{N} \cdot \sum \left(\frac{x_i - \hat{x}_i}{x_i} \right)$)

AAD: Average Absolute Deviation ($AAD = \frac{1}{N} \cdot \sum |x_i - \hat{x}_i|$)

where N is the total number of compounds, x_i is the experimental molar fraction of the compound i and $\hat{x}_i$ is the calculated molar fraction of the compound i obtained by entropy maximization.

whatever the initial database G or G+. However, the obtained molecular compositions are very different depending on the initial database used. To illustrate these differences, comparisons between calculated and experimental molecular compositions of gasoline #A are given in Figure 5. The Root Mean Squared Error (RMSE) is also shown in Figure 5 as an indicator of the quality of both reconstructions. Further statistical comparison of both results can be found in Table 10. With database G, the obtained molecular composition differs significantly from the experimental composition, as shown in Figure 5 and indicated by a RMSE value of 0.84 mol%. In contrast to this, the molecular composition obtained with the database G+ is much closer to the experimental composition *(Fig. 5)* with a RMSE value of 0.25 mol%. This result can easily be explained. Since the initial database G is poorly representative of FCC gasolines in the absence of constraints, it is very difficult for the entropy maximization method to correct this initial bias without major modification of the molar composition of the initial set of compounds. This is due to the fact that the only information available in database G concerns the presence of the compounds, and not their typical abundance. In contrast, database G+ contains not only the information on the type of the compounds that are present, but also on their average mole fraction. Consequently, the molecular composition obtained using database G+ is much more representative of the experimental mole fractions. For this reason, only the reconstructions performed with the database G+ are presented here, and database G has been abandoned for real-life applications.

After reconstruction of the 3 FCC gasolines with the database G+, the results demonstrate a very good agreement between the experimental mole fractions of the 230 identified species obtained by Gas Chromatography and the calculated mole fractions obtained by reconstruction using only the distillation data and overall PIONA analysis as input data. This is illustrated by the various statistical indicators in Table 10 and by the parity plots in Figures 6, 7 and 8, which underline the efficiency of this approach. Indeed, the use of initial database G+ allows to start from the average FCC gasoline and reduces the need for important modifications of the molar fractions of the molecules in the database, which allows to avoid the appearance of a limited number of predominant molecules.

It is also important to note that the use of the simulated distillation in the input data guarantees a Gaussian distribution of the compounds as function of their normal boiling points or their carbon numbers, thereby avoiding the introduction of

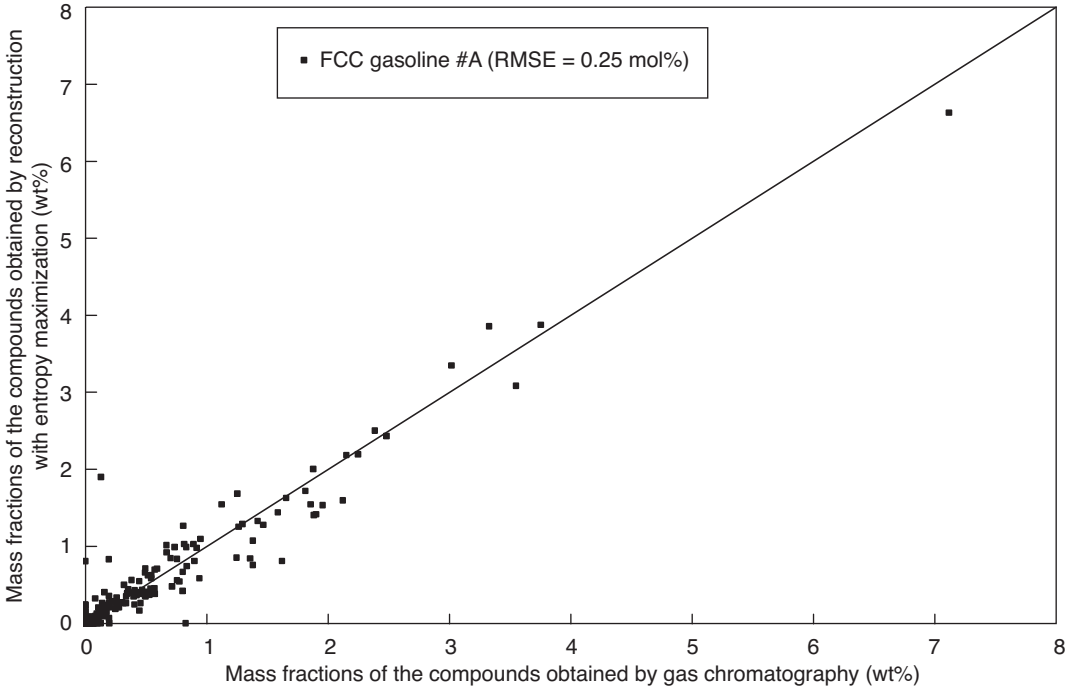

Figure 6

Parity plot of the molecular composition between GC analysis and reconstruction by entropy maximization – Case of the FCC gasoline #A.

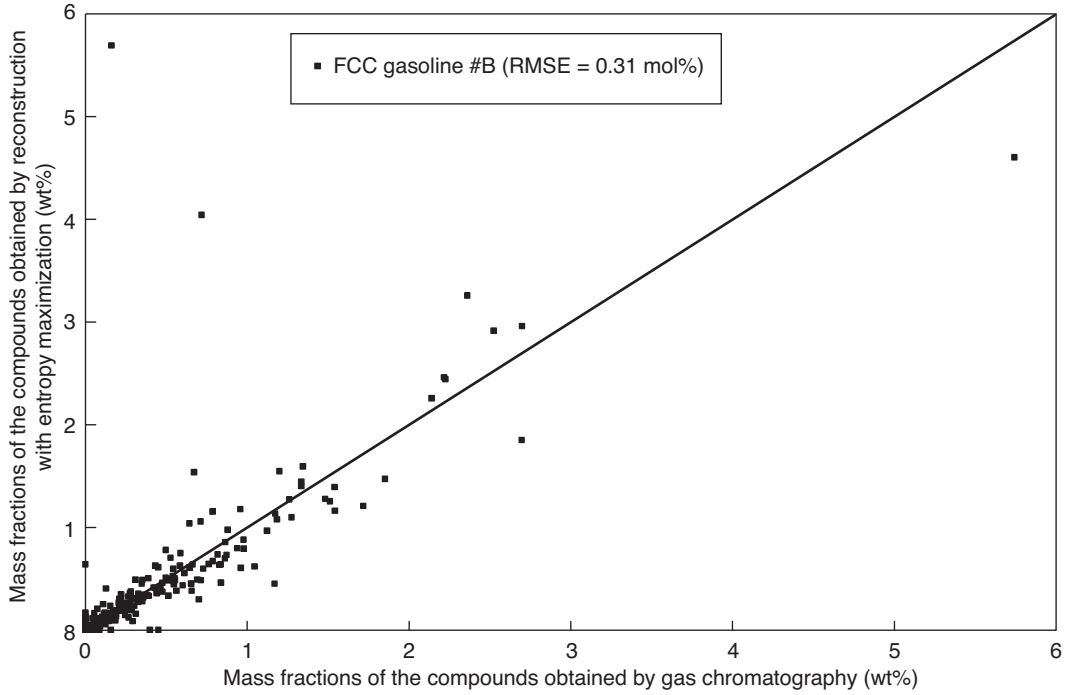

Figure 7

Parity plot of the molecular composition between GC analysis and reconstruction by entropy maximization – Case of the FCC gasoline #B.

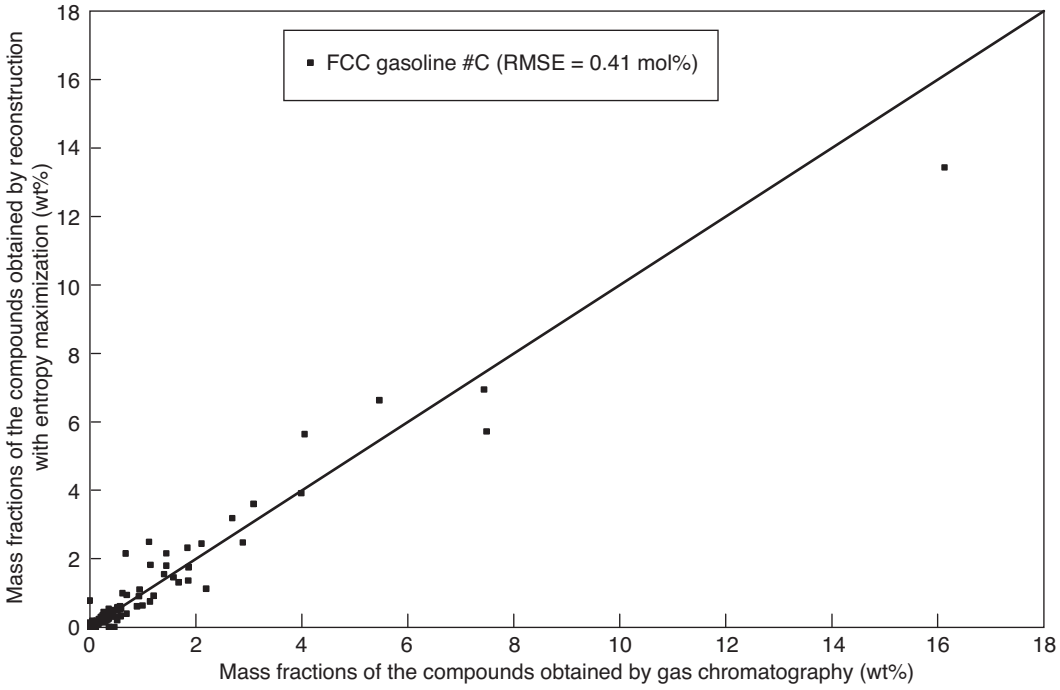

Figure 8

Parity plot of the molecular composition between GC analysis and reconstruction by entropy maximization – Case of the FCC gasoline #C.

additional constraints that force normal distributions *(see Sect. 1.8)* but require additional input data (standard deviation of the Gaussian distribution for each constraint) from the user.

CONCLUSION

The entropy maximization algorithm is an elegant molecular reconstruction method that adjusts the molar fractions of the compounds of a database to verify constraints that are related to the properties of the feedstock to be rebuilt. The theoretical study described in this article allowed to illustrate the following aspects:

- the construction of the initial database of compounds is extremely important because the database must contain compounds that actually exist in the petroleum mixtures. That is why each type of feedstocks should have its own initial database which may depend on the origin of the crude oil and the refinery processes;
- when the initial database is built, it is necessary to translate the mixture properties into constraints. When these constraints are linear or can be transformed into linear constraints, the entropic criterion can be elegantly reorganized, thereby significantly reducing the solution space of the optimization problem, and a semi-algebraic resolution of the entropy maximization problem can be utilized.
- when there are no constraints, the entropy maximization method creates an equimolar mixture, *i.e.* each molar fraction is equal to $1/N$;
- when exact constraints are added, the entropy maximization method modifies the uniform distribution to strictly satisfy these new constraints;
- because the petroleum analyses contain uncertainties, the use of exact constraints is not always ideal. In this case, the entropy maximization method can be modified to handle to constraints with uncertainties. The resolution of the problem is similar to the case with exact constraints;
- when the properties of the initial database are far from the properties of the feedstock to be rebuilt, the entropy maximization method leads to a molar fraction distribution that is characterized by a limited number of predominant molecules, the others having a much lower molar fraction. This is due to the exponential term in the equations that calculate the molar fractions from the Lagrange multipliers. To counter this phenomenon, it is possible to include new additional constraints which force the solution to have Gaussian-type distributions.

The main advantage of molecular reconstruction by entropy maximization resides in the fact that the system with N unknowns (*i.e.*, the molar fractions of the N compounds of the database) is reduced to a system with $J + K + L$ unknowns (*i.e.*, the Lagrange multipliers and normalized residuals associated to the constraints). Depending on the type of petroleum fractions to be reconstructed, N can have an order of magnitude ranging from several hundreds to several tens of thousands. In real case problems, the number of analytical constraints lies typically between 10 and 40. This clearly illustrates that the entropy maximization technique allows to transform an ill-posed optimization problem into a constrained optimization problem with a unique solution, while strongly reducing the computational burden.

Applied to the reconstruction of FCC gasolines, this method was employed to determine the molecular composition of this type of petroleum fractions from partial analyses such as a distillation curve and an overall PIONA analysis. In a first approach, a qualitative database containing 230 different compounds typical of FCC gasolines has been elaborated from detailed gas chromatographic analyses of 15 feedstocks. As this database initially represents an equimolar mixture of the 230 compounds with properties that are very different from those of typical FCC gasolines, the entropy maximization algorithm can not efficiently counter this initial bias and leads to the presence a limited number of dominating molecules. Although the resulting mixtures have the same properties as the FCC gasoline to be reconstructed, their molecular composition is highly unrealistic.

Hence, an alternative database has been developed based on a quantitative approach, in which the same compound is introduced several times in the database in proportion to its molar fraction. The advantage of this quantitative database resides in the fact that the properties of the corresponding equimolar mixture are by construction very close to those of typical FCC gasolines. Applying the entropy maximization method to this quantitative database allows to accurately predict the molecular composition of various FCC gasolines, starting only from their distillation data and overall PIONA analyses. The predicted compositions can be now used to develop molecular-level kinetic models for post-treatment processes, such as selective hydrodesulfurization.

REFERENCES

1 Neurock M. (1992) A Computational Chemical Reaction Engineering Analysis of Complex Heavy Hydrocarbon Reaction Systems, *PhD Thesis*, University of Delaware.

2 Neurock M., Nigam A., Trauth D.M., Klein M.T. (1994) Molecular Representation of Complex Hydrocarbon Feedstocks through Efficient Characterization and Stochastic Algorithms, *Chem. Eng. Sci.* **49**, 24, 4153-4177.

3 Trauth D.M. (1993) Structure of Complex Mixtures through Characterization, Reaction, and Modeling, *PhD Thesis*, University of Delaware.

4 Trauth D.M., Stark S.M., Petti T.F., Neurock M., Klein M.T. (1994) Representation of the Molecular Structure of Petroleum Resid through Characterization and Monte Carlo Modeling, *Energ. Fuel.* **8**, 3, 576-580.

5 Khorasheh F., Khaledi R., Gray M.R. (1998) Computer generation of representative molecules for heavy hydrocarbon mixtures, *Fuel* **77**, 4, 241-253.

6 Hudebine D. (2003) Reconstruction moléculaire de coupes pétrolières, *PhD Thesis*, École Normale Supérieure de Lyon.

7 Hudebine D., Verstraete J.J. (2004) Molecular Reconstruction of LCO Gasoils from Overall Petroleum Analyses, *Chem. Eng. Sci.* **59**, 22-23, 4755-4763.

8 Verstraete J.J., Revellin N., Dulot H., Hudebine D. (2004) Molecular reconstruction of vacuum gasoils, *Prep. Am. Chem. Soc. Div. Fuel Chem.* **49**, 1, 20-21.

9 Liguras D.K., Allen D.T. (1989) Structural Models for Catalytic Cracking 1. Model Compound Reactions, *Ind. Eng. Chem. Res.* **28**, 6, 665-673.

10 Liguras D.K., Allen D.T. (1989) Structural Models for Catalytic Cracking 2. Reactions of Simulated Oil Mixtures, *Ind. Eng. Chem. Res.* **28**, 6, 674-683.

11 Allen D.T., Liguras D.K. (1991) Structural Models of Catalytic Cracking Chemistry: A Case Study of a Group Contribution Approach to Lumped Kinetic Modeling, in *Chemical Reactions in Complex Mixtures: The Mobil Workshop*, Sapre A.V., Krambeck F.J. (eds), Van Nostrand Reinhold, New-York.

12 Quann R.J., Jaff S.B. (1992) Structure-Oriented Lumping: Describing the Chemistry of Complex Hydrocarbon Mixtures, *Ind. Eng. Chem. Res.* **31**, 11, 2483-2497.

13 Quann R.J., Jaffe S.B. (1996) Building Useful Models of Complex Reaction Systems in Petroleum Refining, *Chem. Eng. Sci.* **51**, 10, 1615-1635.

14 Jaffe S.B., Freund H., Olmstead W.N. (2005) Extension of Structure-Oriented Lumping to Vacuum Residua, *Ind. Eng. Chem. Res.* **44**, 9840-9852.

15 Zhang Y. (1999) A Molecular Approach for Characterization and Property Predictions of Petroleum Mixtures with Applications to Refinery Modelling, *PhD Thesis*, University of Manchester.

16 Eckert E., Vanek T. (2005) Extended Utilization of the Characterization of Petroleum Mixtures Based on Real Components, *Chemical Papers* **59**, 6a, 428-433.

17 Eckert E., Vanek T. (2008) Mathematical modelling of selected characterisation procedures for oil fractions, *Chemical Papers* **62**, 1, 26-33.

18 Eckert E., Vanek T. (2009) Improvements in the selection of real components forming a substitute mixture for petroleum fractions, *Chemical Papers* **63**, 4, 399-405.

19 Shannon C.E. (1948) A Mathematical Theory of Communication, *The Bell System Technical J.* **27**, 1, 379-423.

20 Van Geem K.M., Hudebine D., Reyniers M.F., Wahl F., Verstraete J.J., Marin G.B. (2007) Molecular reconstruction of naphtha steam cracking feedstocks based on commercial indices, *Comput. Chem. Eng.* **31**, 9, 1020-1034.

21 Van Geem K.M., Reyniers M.F., Marin G.B. (2008) Challenges of Modeling Steam Cracking of Heavy Feedstocks, *Oil Gas Sci. Technol.* **63**, 1, 79-94.

22 Song C. (2003) An overview of new approaches to deep desulfurization for ultra-clean gasoline, diesel fuel and jet fuel, *Catal. Today* **88**, 1-4, 211-263.

23 Thermodynamics Research Center (1998) TRC Thermodynamic Tables, *The Texas A&M University System*, College Station, TX (USA).

Final manuscript received in April 2011
Published online in June 2011

ANNEX A

The aim of this annex is to propose to the reader the new group contribution correlations [6] that were developed to estimate the normal boiling point and the density at 20°C of different compounds from their 2D structure.

A.1 Description of the Compound Databases Used to Develop the Group Contribution Correlations

In order to calibrate the group contribution correlations for the normal boiling point and the density at 20°C, an initial database of pure compounds has been created from the compiled TRC database [23]. At the time of writing, the database contains 2 569 hydrocarbon and sulfur compounds but can be easily extended by including nitrogen and oxygen compounds if required. For each compound, the database contains the following properties:

- the boiling point for a given pressure. In the majority cases, the pressure is equal to 1 atm, but for some compounds, the boiling point has been determined under reduced pressure;
- the liquid density at 2 temperatures. In the majority cases, the first temperature is 20°C while the second temperature is 25°C. However, some compounds are solid at these temperatures and the density is then given at the saturation temperature.

After removing the compounds whose boiling point is not provided at atmospheric pressure, the initial database is reduced to 1 827 compounds. This new database corresponds to the calibration database which will be used to develop the correlation for the normal boiling point T_b. Table A.1 gives an overview of this database in terms of number of chemical families and atom number families.

When the initial database is filtered to keep compounds whose density is at 20°C, the number of compounds is equal to 1 845. This new database whose main characteristics are also given in Table A.1 is then used to determine the correlation for the density at 20°C.

TABLE A.1

Number of compounds contained in the different databases by chemical family and by total number of atoms (except hydrogen)

Chemical families	Calibration database for T_b	Calibration database for d^{20}
Paraffins	727	726
Olefins	195	195
Diolefins	29	29
Alkynes	83	83
Mononaphthenes	174	174
Dinaphthenes	16	10
Trinaphthenes	4	0
Monoaromatics	118	118
Diaromatics	70	46
Triaromatics	13	3
Tetraaromatics	8	4
Mononaphtheno-monoaromatics	85	158
Mononaphtheno-diaromatics	3	2
Dinaphtheno-monoaromatics	2	2
Bicyclohexyls	16	16
Biphenyls	63	49
Cycloolefins	42	42
Thiols	48	45
Sulfides	61	58
Thiophenes	18	17
Others	52	68
Total	1 827	1 845
Total number of atoms (except hydrogen)	**Calibration database for T_b**	**Calibration database for d^{20}**
Hydrocarbon compounds with 5 to 12 carbon atoms	1 237	1 239
Hydrocarbon compounds with 13 to 20 carbon atoms	309	329
Hydrocarbon compounds with 21 to 42 carbon atoms	154	157
Sulfur compounds with 5 to 12 atoms (C + S)	95	88
Sulfur compounds with 13 to 20 atoms (C + S)	28	28
Sulfur compounds with 21 to 42 atoms (C + S)	4	4
Total	1 827	1 845

A.2 Description of the Group Contribution Correlations Used to Calculate the Normal Boiling Point and the Density at 20°C

Group contribution methods are a set of more or less complex correlations that allow to calculate the properties of pure compounds from their chemical structure. They are based on the principle that a molecule can be broken down into a certain number of elementary chemical groups which may be either single atoms (aromatic carbon, naphthenic carbon), pairs of atoms (olefinic, cyanide), or larger functional groups (carboxyl, amide). Each group, whatever its place in the molecule, has its own contribution that adds to the properties of the molecule. The general equation used to calculate a property P using a group contribution method is therefore as follows:

$$P = f\left(\sum_i n_i \cdot C_i\right) \tag{A.1}$$

where P Property to be determined

$f(\ )$ Relation between the property and the group contributions

n_i Number of groups of type i

C_i Contribution of a group of type i

Applied to the case of the prediction of the normal boiling point, Equation (A.1) used in this work has the following form:

$$\exp\left(T_b / T_{b,0}\right) = \sum_i n_i \cdot \Delta T_{b,i} \qquad \text{(A.2)}$$

where T_b Normal boiling point of the pure compound (K)

$T_{b,0}$ Reference normal boiling point (K)

n_i Number of groups of type i

$\Delta T_{b,i}$ Contribution for T_b of a group of type i

The density at 20°C is predicted by means of Equation (A.3) and Equation (A.4):

$$d^{20} = \frac{M}{V_m^{20}} \qquad \text{(A.3)}$$

$$V_m^{20} = \sum_i n_i \cdot \Delta V_{m,i}^{20} \qquad \text{(A.4)}$$

with d^{20} Density at 20°C of the pure compound (g/cm^3)

M Molecular weight of the pure compound (g/mol)

TABLE A.2

Contributions of the different structural groups for the calculation of the normal boiling point and the density at 20°C

	Structural groups	Calculation of T_b [(a)]		Calculation of d^{20}	
		Number[(b)]	Contribution	Number[(b)]	Contribution
$-CH_3$	aliphatic	*5 333*	0.87577	*5 520*	32.14
$-CH_2-$	aliphatic	*8 617*	0.31012	*8 842*	16.38
>CH–	aliphatic	*1 325*	–0.33431	*1 346*	–0.93
>C<	aliphatic	*482*	–0.92606	*491*	–19.45
$-CH_2-$	naphthenic	*1 319*	0.38519	*1 333*	13.93
>CH–	naphthenic substituted	*395*	–0.25190	*438*	–1.16
>CH–	naphthenic condensed	*42*	–0.13434	*20*	–3.98
>C<	naphthenic disubstituted	*44*	–0.98407	*92*	–15.89
>C<	naphthenic substituted and condensed	*6*	–0.82463	*4*	–18.28
$=CH_2$	olefinic	*141*	0.86394	*141*	29.70
=CH–	olefinic	*281*	0.32322	*288*	13.55
=C<	olefinic	*120*	–0.25744	*117*	–4.12
=C=	olefinic	*8*	0.40947	*8*	8.79
=CH–	cycloolefinic	*72*	0.37022	*77*	10.97
=C<	cycloolefinic condensed	*0*	–0.14931	*0*	–6.94
=C<	cycloolefinic substituted	*34*	–0.22548	*49*	–4.03
CH	triple bound	*45*	0.80414	*44*	25.79
C–	triple bound	*109*	0.38502	*114*	8.85
=CH–	aromatic	*2 280*	0.38136	*2 166*	11.22
=C<	aromatic substituted	*714*	–0.14943	*678*	–6.15
=C<	aromatic condensed peripherical	*420*	0.00673	*476*	–7.74
=C<	aromatic condensed internal	*8*	–0.39948	*2*	–10.97
–SH	thiophenolic	*9*	1.51910	*6*	27.55
–SH	aliphatic or naphthenic	*36*	1.53226	*37*	28.67
–S–	aliphatic	*32*	0.92809	*28*	12.53
–S–	thiophenic	*17*	0.71371	*16*	12.46
–S–	naphthenic	*8*	1.01684	*9*	9.81
–S–	benzoïc	*8*	0.87408	*9*	12.34
–S–	disulfide	*16*	0.74243	*16*	14.53
	Molecule with 1 ring	*383*	0.89744	*383*	25.44
	Molecule with 2 rings	*267*	1.76513	*301*	52.10
	Molecule with 3 rings	*19*	3.10092	*7*	73.74
	Molecule with 4 rings	*8*	5.62178	*4*	100.00

(a) Group contributions in relation to $T_{b,0}$ = 307.63 K.

(b) Total number of structural groups present in the calibration database.

V^{20}_m Molar volume of the pure compound at 20°C (cm^3/mol)

$\Delta V^{20}_{m,i}$ Contribution for d^{20} of a group of type i (cm^3/mol)

After the multi-linear regression carried out on the calibration databases for T_b and d^{20}, the resulting values of the contributions for the various structural groups for both properties are given in Table A.2. The total number of the structural groups in the both calibration databases is also indicated.

A.3 Validation of the Developed Group Contribution Correlations

After optimization, the statistical properties for the group contribution method for the normal boiling point are:

– average absolute relative error: 1.20%;

– average absolute deviation: 5.7 K;

– standard deviation: 7.4 K.

The parity plot comparing the experimental normal boiling points and those calculated by the group contribution method is given in Figure A.1. The corresponding error distribution is illustrated in Figure A.2.

For the prediction of the density at 20°C, the statistical data after optimization are:

– average absolute relative error: 1.13%;

– average absolute deviation: 0.0090 g/cm^3;

– standard deviation: 0.0129 g/cm^3.

The parity plot comparing the experimental densities at 20°C and those calculated by the group contribution method is given in Figure A.3. The corresponding error distribution is illustrated in Figure A.4.

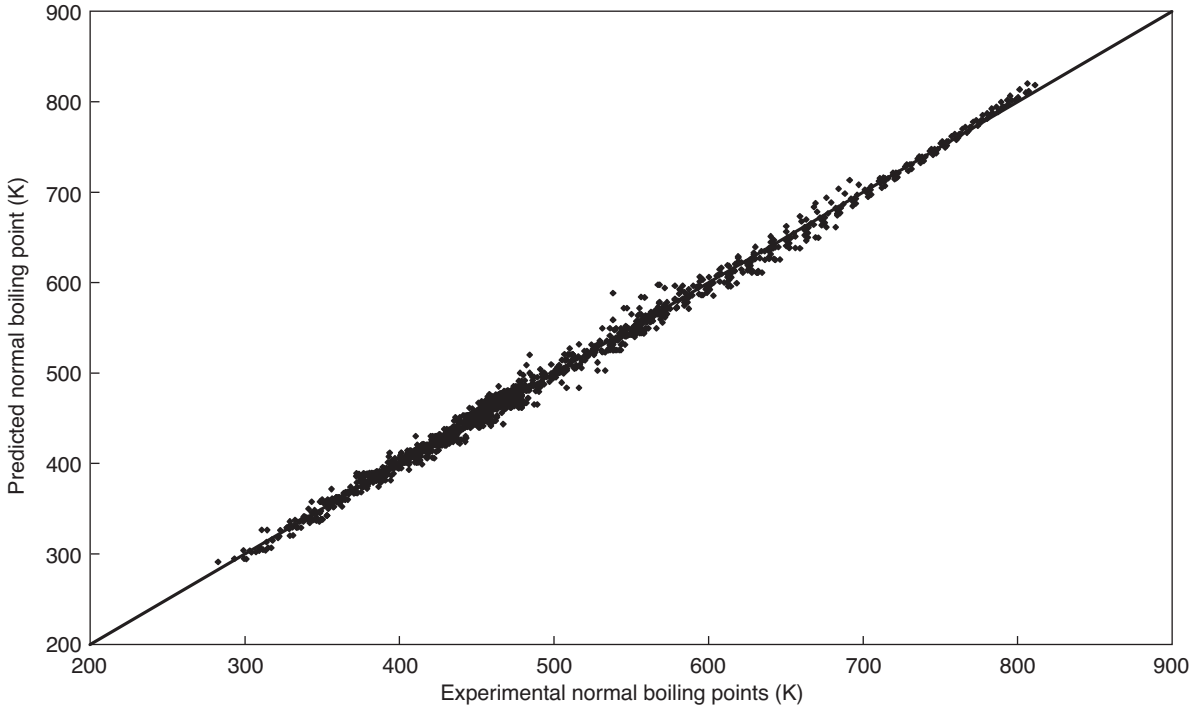

Figure A.1

Parity plot for the prediction of the normal boiling point.

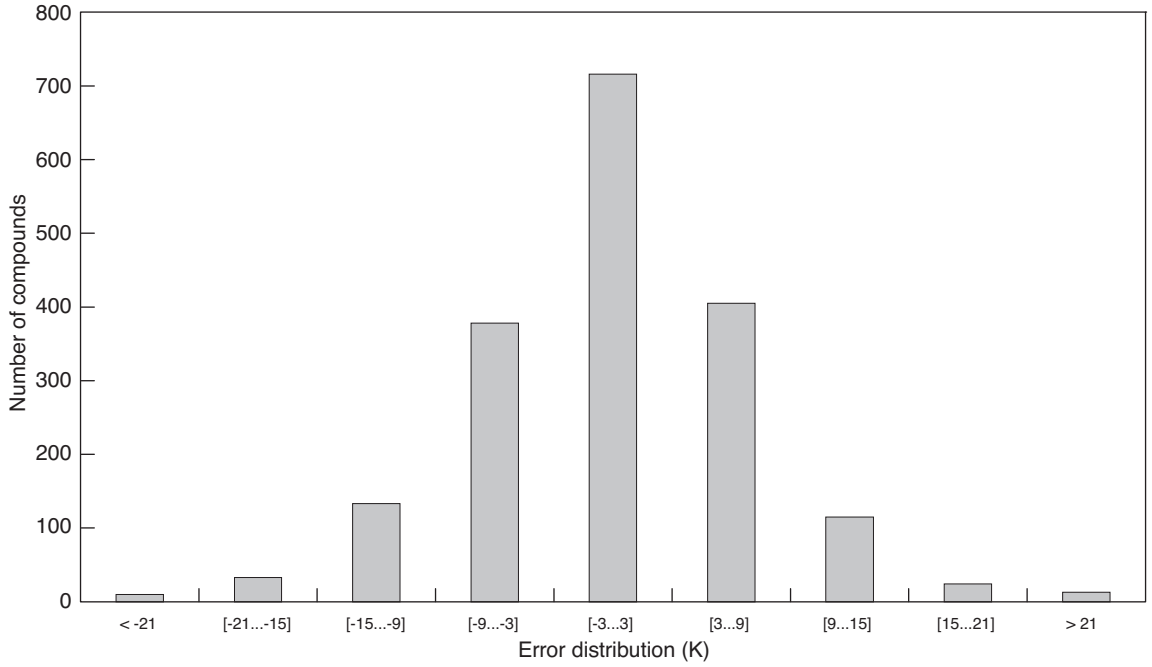

Figure A.2

Error distribution for the prediction of the normal boiling point.

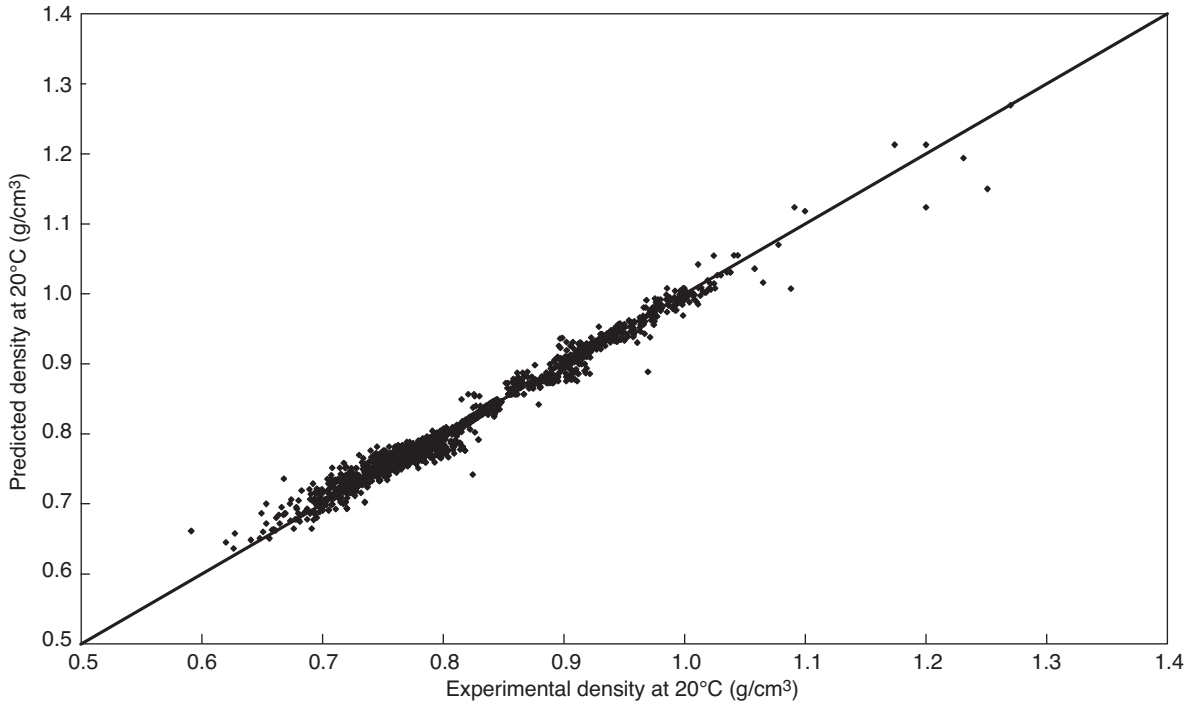

Figure A.3

Parity plot for the prediction of the density at 20°C.

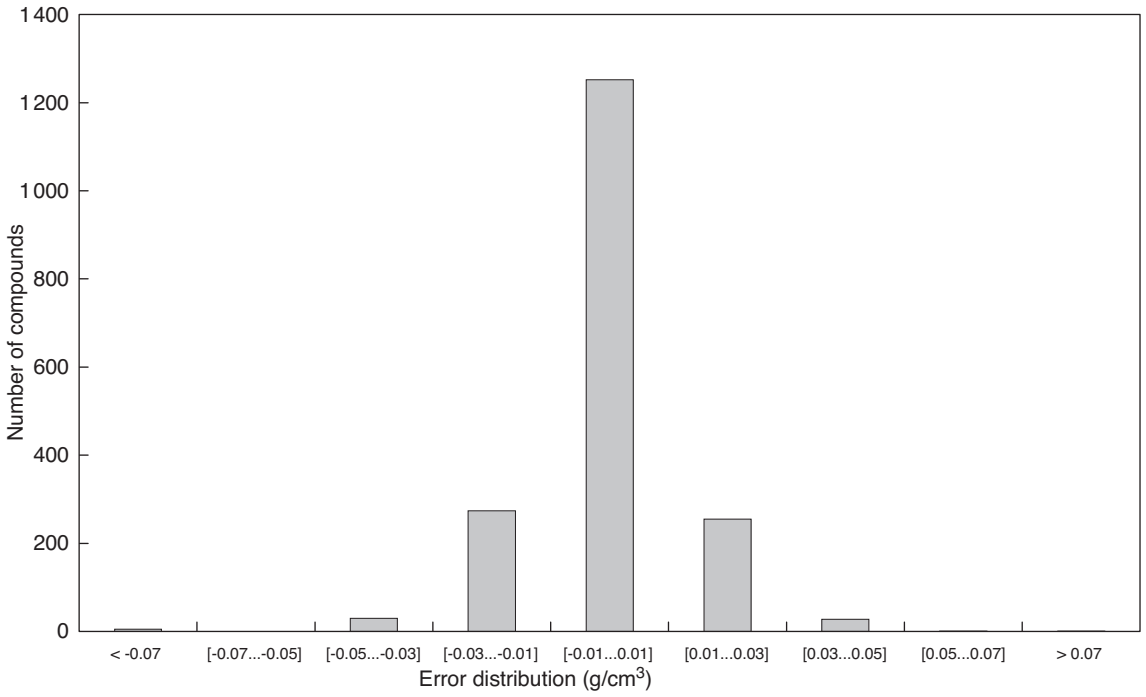

Figure A.4

Error distribution for the prediction of the density at 20°C.