



This paper is a part of the hereunder thematic dossier published in OGST Journal, Vol. 68, No. 2, pp. 187-396 and available online [here](#)

Cet article fait partie du dossier thématique ci-dessous publié dans la revue OGST, Vol. 68, n°2, pp. 187-396 et téléchargeable [ici](#)

DOSSIER Edited by/Sous la direction de : Jean-Charles de Hemptinne

InMoTher 2012: Industrial Use of Molecular Thermodynamics InMoTher 2012 : Application industrielle de la thermodynamique moléculaire

Oil & Gas Science and Technology – Rev. IFP Energies nouvelles, Vol. 68 (2013), No. 2, pp. 187-396

Copyright © 2013, IFP Energies nouvelles

- 187 > Editorial
- 217 > *Improving the Modeling of Hydrogen Solubility in Heavy Oil Cuts Using an Augmented Grayson Streed (AGS) Approach*
Modélisation améliorée de la solubilité de l'hydrogène dans des coupes lourdes par l'approche de *Grayson Streed Augmenté* (GSA)
R. Torres, J.-C. de Hemptinne and I. Machin
- 235 > *Improving Group Contribution Methods by Distance Weighting*
Amélioration de la méthode de contribution du groupe en pondérant la distance du groupe
A. Zaitseva and V. Alopaeus
- 249 > *Numerical Investigation of an Absorption-Diffusion Cooling Machine Using C_3H_8/C_9H_{20} as Binary Working Fluid*
Étude numérique d'une machine frigorifique à absorption-diffusion utilisant le couple C_3H_8/C_9H_{20}
H. Dardour, P. Cézac, J.-M. Reneaume, M. Bourouis and A. Bellagi
- 255 > *Thermodynamic Properties of 1:1 Salt Aqueous Solutions with the Electrolytic Equation of State*
Propriétés thermophysiques des solutions aqueuses de sels 1:1 avec l'équation d'état de réseau pour électrolytes
A. Zuber, R.F. Checoni, R. Mathew, J.P.L. Santos, F.W. Tavares and M. Castier
- 271 > *Influence of the Periodic Boundary Conditions on the Fluid Structure and on the Thermodynamic Properties Computed from the Molecular Simulations*
Influence des conditions périodiques sur la structure et sur les propriétés thermodynamiques calculées à partir des simulations moléculaires
J. Janeček
- 281 > *Comparison of Predicted pKa Values for Some Amino-Acids, Dipeptides and Tripeptides, Using COSMO-RS, ChemAxon and ACD/Labs Methods*
Comparaison des valeurs de pKa de quelques acides aminés, dipeptides et tripeptides, prédites en utilisant les méthodes COSMO-RS, ChemAxon et ACD/Labs
O. Toure, C.-G. Dussap and A. Lebert
- 299 > *Isotherms of Fluids in Native and Defective Zeolite and Alumino-Phosphate Crystals: Monte-Carlo Simulations with "On-the-Fly" ab initio Electrostatic Potential*
Isothermes d'adsorption de fluides dans des zéolithes silicées et dans des cristaux alumino-phosphatés : simulations de Monte-Carlo utilisant un potentiel électrostatique *ab initio*
X. Rozanska, P. Ungerer, B. Leblanc and M. Yiannourakou
- 309 > *Improving Molecular Simulation Models of Adsorption in Porous Materials: Interdependence between Domains*
Amélioration des modèles d'adsorption dans les milieux poreux par simulation moléculaire : interdépendance entre les domaines
J. Puibasset
- 319 > *Performance Analysis of Compositional and Modified Black-Oil Models For a Gas Lift Process*
Analyse des performances de modèles black-oil pour le procédé d'extraction par injection de gaz
M. Mahmudi and M. Taghi Sadeghi
- 331 > *Compositional Description of Three-Phase Flow Model in a Gas-Lifted Well with High Water-Cut*
Description de la composition des trois phases du modèle de flux dans un puits utilisant la poussée de gaz avec des proportions d'eau élevées
M. Mahmudi and M. Taghi Sadeghi
- 341 > *Energy Equation Derivation of the Oil-Gas Flow in Pipelines*
Dérivation de l'équation d'énergie de l'écoulement huile-gaz dans des pipelines
J.M. Duan, W. Wang, Y. Zhang, L.J. Zheng, H.S. Liu and J. Gong
- 355 > *The Effect of Hydrogen Sulfide Concentration on Gel as Water Shutoff Agent*
Effet de la concentration en sulfure d'hydrogène sur un gel utilisé en tant qu'agent de traitement des venues d'eaux
Q. You, L. Mu, Y. Wang and F. Zhao
- 363 > *Geology and Petroleum Systems of the Offshore Benin Basin (Benin)*
Géologie et système pétrolier du bassin offshore du Benin (Benin)
C. Kaki, G.A.F. d'Almeida, N. Yalo and S. Amelina
- 383 > *Geopressure and Trap Integrity Predictions from 3-D Seismic Data: Case Study of the Greater Ughelli Depobelt, Niger Delta*
Pressions de pores et prévisions de l'intégrité des couvertures à partir de données sismiques 3D : le cas du grand sous-bassin d'Ughelli, Delta du Niger
A.I. Opara, K.M. Onuoha, C. Anowai, N.N. Onu and R.O. Mbach

Improving Group Contribution Methods by Distance Weighting

A. Zaitseva* and V. Alopæus

Aalto University, School of Chemical Technology, Department of Biotechnology and Chemical Technology, P.O.B. 16100, 00076 Aalto - Finland
e-mail: anna.zaitseva@aalto.fi - ville.alopæus@aalto.fi

* Corresponding author

Résumé — Amélioration de la méthode de contribution du groupe en pondérant la distance du groupe — La nouvelle approche pour améliorer les estimations existantes de la méthode de contribution du groupe est développée. Au lieu d'utiliser les contributions fixes, les contributions pondérées sont optimisées pour chaque propriété de la base de données. Les facteurs de pondération sont calculés en se basant sur les similitudes entre le composé, dont les propriétés sont estimées, et les autres composés de la base de données. En utilisant cette approche, les substances qui sont chimiquement plus proches des substances en question sont systématiquement plus pondérées. Ainsi la nature générale de l'approche est maintenue.

La nouvelle technique est appliquée aux méthodes de contribution du groupe de Joback et Reid (1987), ainsi que de Marrero et Gani (2001). Les performances sont démontrées en utilisant la prédiction du point normal d'ébullition (NBP, Normal Boiling Point). La moyenne de l'erreur absolue des prédictions de NBP est réduite de 6,3 K pour la méthode de Joback et Reid, et de 4 K pour la méthode de Marrero et Gani. Les autres propriétés physiques comme la pression critique, le volume, l'énergie de formation, la température de fusion et l'enthalpie de fusion sont estimées en appliquant la nouvelle technique à la méthode de Joback et Reid.

Abstract — Improving Group Contribution Methods by Distance Weighting — A novel approach for improving estimations of existing group contribution methods is developed. Instead of fixed contributions, weighted contributions are optimized for each property estimation using a database. These weighting factors are calculated based on similarities between the compound whose properties are estimated and the other compounds in the database. By this approach, those components which are chemically more similar to the estimated one can systematically be given more weight in the estimation process, while sustaining the general nature of the group contribution methods.

The new approach was applied to the Joback and Reid (1987), and the Marrero and Gani (2001) group contribution methods. The performances were demonstrated on Normal Boiling Point (NBP) predictions. The absolute average error of NBP estimations was reduced by 6.3 K for the Joback and Reid method and by 4 K for the Marrero and Gani method. Other physical properties such as critical pressure, volume, formation energies, melting temperature and fusion enthalpy were estimated by applying the new technique to Joback and Reid method.

NOMENCLATURE

Abbreviations

AAE	Absolute Average Error $AAE = \frac{1}{m} \sum_{j=1}^m P_j^{est} - P_j^{meas} $
AE	Absolute Error - $ P_j^{est} - P_j^{meas} $
GC	Group Contribution
GCM	Group Contribution Method
JR	Joback Reid method
MG	Marrero and Gani method
NBP	Normal Boiling Point
NW	Uniform Weighting technique
DW	Distance Weighting technique

Notations

a_0	Group contribution method constant
a_i	Group contribution coefficient of group i
(a)	Column matrix of the group contributions
d_j	Distance function for component j
n_i	Number of functional groups of type i
$[n]$	Matrix of the number of groups
N_G	Total number of groups in the method
m	Total number of compounds in the database
P	Pure component property
(P_{meas})	Column matrix of measured pure component properties
$[W]$	Matrix of weights
ε	Constant used in calculation of weights, equal to 1 in this work

Sub- and superscripts

i	Group index
j	Compound index
$meas$	Measured
est	Estimated

INTRODUCTION

Pure component properties are the most important parameters needed for thermodynamic calculations in chemical engineering. Experimental measurement of these properties can be difficult, time consuming or even impossible in case of toxic or degrading compounds. Therefore, predictive techniques are a valuable alternative to experimental measurements.

Predictive methods estimate the desired property of a chemical compound based on the experimental values of the same property measured for other compounds, which are in some way similar to the compound of interest. Usually, the molecule of interest is represented as a combination of groups and the contribution of each group is used in estimating the properties of the whole molecule. In most cases, the groups are combinations of atoms (*e.g.* in the method of Joback and Reid (JR) (Joback and Reid, 1987), in the methods by Marrero and Gani (MG) and Nannoolal *et al.* (Marrero and Gani, 2001; Nannoolal *et al.*, 2004)), but they can even represent bonds in the molecule (*e.g.* in the method by Marrero-Morejon and Pardillo-Fontdevila (Marrero-Morejon *et al.*, 1999; Marrero-Morejón and Pardillo-Fontdevila, 2000)). The group contributions are either summed up, or their sum is used in more complex correlations (Constantinou and Gani, 1994; Nannoolal *et al.*, 2004; Wang *et al.*, 2009). Each group contribution is typically assumed independent of the contributions of the other groups. Group Contribution Methods (GCMs) can account for the relative position of the groups in a molecule by introducing second and third order groups (cross – effects), use of different topological indexes (Gani *et al.*, 2005; González *et al.*, 2007; Wakeham *et al.*, 2002) or by applying proximity parameters within the group contribution summation (Vijande *et al.*, 2010). The methods by Marrero and Gani and Nannoolal *et al.* (Marrero and Gani, 2001; Nannoolal *et al.*, 2004) are recently published examples of GCMs for estimation of pure compound physical properties. Another wide area of application of the GCMs is the prediction of a model parameters for description of chemical mixture behavior (Gmehling, 2009; Mustafa *et al.*, 2011; Peters *et al.*, 2012; Soria *et al.*, 2011; Tochigi *et al.*, 2002; Vijande *et al.*, 2010).

Even though GCMs are widely used in chemical engineering calculations, their accuracy is uncertain. The error of a physical property estimation can vary greatly among different chemical classes (Rowley *et al.*, 2007), since the contribution of the various groups is averaged over compounds that are chemically rather different. The error of the estimation methods may be reduced by considering that similar compounds often behave similarly. In a specific class of chemicals, a property can be estimated by averaging the property values. However, for complex molecules with rare functional groups or group combinations, a property averaging over neighbor compounds would not account for all the dependencies of the property on the molecular structure.

In this work, we developed a novel calculation procedure that is applied to existing GC methods in order to exploit the concept of chemical similarity and improve their estimations. The proposed estimation technique can be applied to any GCM. The advantages of this methodology are presented here using the JR method because of its simple form and wide-spread use. We also tested the new technique on the more advanced method by Marrero and Gani.

1 OPTIMIZATION OF GROUP CONTRIBUTIONS

Group Contribution Methods use the principle that a certain structural group in a molecule influences physical properties always in the same way. Thus, for component j the physical property $P_{calc,j}$ is expressed as a function of the group contributions, *i.e.* $P_{calc,j} = f\left(\sum_{i=0}^{N_G} a_i n_i\right)$ where a_i is the group contribution coefficient of group i , $n_{i,j}$ is the number of functional groups of type i and i varies from 0 to the total number of groups N_G . Based on a physical property database, the group contributions are estimated by minimizing the sum of the squared differences between calculated and measured properties:

$$\min \sum_j (P_{calc,j} - P_{meas,j})^2 \quad (1)$$

In this work, linear or linearized (see examples in *Tab. 1*) representations of the pure component properties were used:

$$P_{calc} = a_0 + \sum_{i=1}^{N_G} a_i n_i \quad (2)$$

TABLE 1

Summary of NBP estimation GCMs used in this work

	Joback and Reid (JR) (1987)	Marrero and Gani (MG) (2001)
Method order	Only 1 st order groups are used	1 st , 2 nd , 3 rd order groups are used
NBP equation	$T_b = T_{b0} + \sum_{i=1}^{N_G} a_i n_i$	$T_b = T_{b0} \cdot \text{LN}\left(\sum_{i=1}^{N_G} a_i n_i\right)$
T_{b0} , original	198.2	222.543
Linearized property for GC optimization	$T_b = \sum_{i=0}^{N_G} a_i n_i$ where $n_0 = 1$	$\exp\left(\frac{T_b}{T_{b0}}\right) = \sum_{i=1}^{N_G} a_i n_i$ where $T_{b0} = 222.543$
Optimization procedure	One step least square minimization in matrix form (Eq. 5, 9)	1 st step – Eq. (5) or (9) for the first order groups; 2 nd step – Eq. (6) or (10)

In the ideal case, the calculated properties (P_{calc}) should be equal to measured property (P_{meas}):

$$P_{meas,j} = P_{calc,j} = a_{0,j} + \sum_{i=1}^{N_G} a_i n_{i,j} \quad (3)$$

and thus in matrix form Equation (3) becomes:

$$(P_{meas}) = [n](a) \quad (4)$$

(P_{meas}) is the column matrix of the measured properties, $[n]$ is the matrix of the number of groups and (a) is the column matrix of the group contributions. The number of rows in each

matrix is the number of components whose measured properties are available. The number of columns in matrix $[n]$ is the number of groups plus one. The first column in matrix $[n]$ is solely filled with ones. Thus, the first element of (a) represents a constant a_0 . The other elements of (a) are the contribution coefficients of the groups according to Equation (2).

In this work, the distance weighted optimization of group contributions was compared with the results of traditional non-weighted optimization.

1.1 Non-Weighted Optimization

In non-weighted optimization group contribution coefficients were solved by means of the matrix inversion, according to Equation (5):

$$(a) = \left[[n]^T [n] \right]^{-1} [n]^T (P_{meas}) \quad (5)$$

In more advanced GCMs second and third order groups are employed, which account for functional group proximity (Constantinou and Gani, 1994; Cordes and Rarey, 2002; Marrero and Gani, 2001). In our work second level GC ($a^{(2)}$) were optimized by minimizing the difference between measured properties and the properties estimated using the lower order group contributions (Eq. 5). Note that (a) and ($a^{(2)}$) are independent vectors and matrix $[n]$ is different for Equation (5) – $[n]$ and for Equation (6) – $[n^{(2)}]$:

$$(a^{(2)}) = \left[[n^{(2)}]^T [n^{(2)}] \right]^{-1} [n^{(2)}]^T (P_{meas} - P_{calc}(a)) \quad (6)$$

This two stage optimization technique was used for comparative purposes, *i.e.* because it was utilized in the mentioned above multilevel GCMs. In accordance with our preliminary investigations, simultaneous optimization of (a) and ($a^{(2)}$) yields a better fit of the measured properties but changes the meaning of the second order groups that are to be only corrections for estimations made with the first order group contributions.

1.2 Distance Weighted Optimization

Utilization of weighting can significantly improve the accuracy of the optimization methods (Dudani, 1976). In the optimization of group contributions, weights can be used to exploit the structural similarities among compounds. Structural groups of GCMs are a natural way of taking into account structural similarities and they were used for calculating the distance function (d_j) in accordance with Equation (7).

$$d_j = \sum_{i=1}^{N_G} (n_{i,j} - n_{i,est})^2 + \left(\sum_{i=1}^{N_G} n_{i,j} - \sum_{i=1}^{N_G} n_{i,est} \right)^2 \quad (7)$$

where i is the group index, j is the index of the compound in the database and the subscript est indicates the compound under estimation. Other definitions of the distance function are also possible, depending on the property under estimation. High influence of some groups on the property can be emphasized by non equal contribution of the different groups into the first sum of Equation (7). The second term in Equation (7) slightly improves the NBP estimations by emphasizing the difference in size among the molecules. The weights were obtained from the distance function by combining Equations (7) and (8).

$$W_j = \frac{1}{d_j + \varepsilon} \quad (8)$$

The factor ε is used to prevent over-emphasizing of very similar compounds and to avoid division by zero, when the selected groups cannot discriminate between two different chemical compounds (*e.g.* stereo-isomers). In this work, ε was given the value of 1. It is clear from Equation (8) that weights are larger for the components, which are chemically similar to the estimated one, *i.e.* with a small distance.

The Group Contributions were optimized using the least square minimizations as shown in Equations (9) and (10). It is important to emphasize that the optimization is made for each estimation task separately because the weights are specific to the compound under estimation:

$$(a) = \left[[n]^T [W] [n] \right]^{-1} [n]^T [W] (P_{meas}) \quad (9)$$

$$(a^{(2)}) = \left[[n^{(2)}]^T [W] [n^{(2)}] \right]^{-1} [n^{(2)}]^T [W] (P_{meas} - P_{calc}) \quad (10)$$

where $[W]$ is a matrix containing the weights on its diagonal and zeros elsewhere.

Alternatively to the computational procedure described, other optimization approaches can be used. Various weighting strategies are also possible. For example, weights can be calculated based on dipole moment closeness or larger weights can be given for compounds belonging to the same chemical class. Uncertainties of experimental data can be taken into account in the weighting procedure, as in Kang *et al.* (2011). With respect to the dependence of physical properties on the Group Contributions other relations than the linear can be used. Non-linear property dependence will require a different minimization approach.

1.3 Estimation of Physical Property with Distance Weighted GCM

For estimation of a physical property a database (DB) of compounds with the measured physical property should be available. Each compound of the DB has to be divided into groups as determined in a GCM that was taken as a basis for the estimation. Then, weights of each compound of the

database are estimated as described in Section 1.2. Group Contribution values are optimized in accordance with Equations (9) and (10). After that, the physical property is calculated in accordance with Equation (2). See also examples of the NBP estimation in the Appendix. Though the procedure of the property estimation thus requires optimization of the GC values at each estimation task, improvement of accuracy of the basic method is notable. Moreover, the distance weighting method can be easily implemented into most of modern flowsheet simulators, because they have already their own compound databases and the GC value optimization takes less than a second on a Pentium computer.

2 PREDICTION OF NORMAL BOILING TEMPERATURE

2.1 Database

In our method, a physical property database is required for optimizing the Group Contributions and for estimating unknown physical properties. Most modern process flowsheet simulators and estimation packages exploit extensive sets of known physical properties and these databases are their essential part. The choice of the compounds included in the database depends on the availability of experimentally measured physical properties. For this work, the Normal Boiling Point (NBP) was chosen because of its importance as a pure component property and because experimental values of the NBP are available for many compounds. The experimental data included in the database were taken from the handbook by Poling *et al.* (2001). Although this set of components and experimental data is much more restricted than the databases used in commercial simulators, it was chosen because it is publicly available. More extensive databases are expected to result in more accurate estimations both in the traditional group contribution methods and in the presented weighted group contribution approach. The database by Poling *et al.* (2001) was nevertheless found to be extensive enough for testing the effect of the group contribution weighting.

In our database, the chemical compounds are divided into structural groups. The division of the compounds into predetermined groups can be automated (Kolská and Petrus, 2010; Raymond and Rogers, 1999). Components whose NBP has neither been measured experimentally, nor calculated based on experimental values were excluded. In this way, the values estimated with our method do not incorporate errors from other estimation methods. In order to determine the contribution of a group in a reasonably reliable manner, the database should include a certain number of compounds containing that group. The minimum amount of the compounds was chosen to be three, *i.e.* if only two compounds contain some particular group, those compounds were excluded from the database and that group was not considered in the analysis.

The total number of components in our database was 386, when using the Joback method. From the components available in the handbook by Poling *et al.* (2001) only the

components that can be divided into Joback groups were selected for the database. In addition to the NBP from the Poling database, the NBP of dimethylacetylene was added to our database from DIPPR Project 801 (2009). Dimethylacetylene was included for the optimization of the triple bond group ($-C\#$) in order to avoid singularity of the group matrix $[n]$. This singularity occurs, if at least two columns become linearly dependent. For example, if in our database all alkynes were represented only by 1-alkynes the $CH\#$ and $-C\#$ group columns would become linearly dependent, therefore, the system of equations represented by Equation (3) could not be solved.

The estimation of the NBP based on the MG method was made for selected classes of compounds, *i.e.* alkanes and alcohols. 40 alcohols and 67 alkanes from the Poling handbook were included in our database. This restriction of the database was made to reduce the number of treated groups in order to improve the representation of each group. In the Poling database only the classes of alkanes and alcohols were large enough for reliable optimization of second order group contributions of the MG method.

Naturally, such restriction does not affect the general objective of this paper, *i.e.* to show the applicability of the proposed technique and the benefits of distance weighting optimization.

2.2 Accuracy of the Tested Models

Our new method for the optimization of group contributions utilizes the form of the property dependence on the group contribution and the group definition of available GCMs. Two GC models were used for the demonstration of our approach: Joback and Reid (JR) and Marrero and Gani (MG) (see *Tab. 1*).

The JR method is a one-level GC method, *i.e.* only 1st order structural groups are used for calculating physical properties. In this method, NBP is calculated as a linear sum of group contributions with the additional compound independent constant. The Marrero and Gani method includes three levels (1st, 2nd and 3rd order groups) and utilizes an exponential function for the estimation of NBP. To be able to use the matrix calculations (*Eq. 5, 6, 9, 10*), $\ln(T_b/T_{b0})$ were utilized as a linearized property (*Tab. 1*), where the constant T_{b0} was taken from the original MG method. Only 1st and 2nd order groups were used from the MG method. Reliable optimization of the third order group contributions was impossible using our database (see *Sect. 2.1 Database*), since only a few compounds in the database contain the third order groups.

For error analysis of the proposed methods estimation of the NBP was made one by one for every compound in the database. For each estimation task, the estimated component was excluded from the database and its NBP was calculated as described in Section 1.3. The optimized contributions (*a*)

used in each estimation task were unique and each contribution set was only used once for the estimation of NBP of a single compound, in accordance to Equation (2).

In order to see the effect of the weights on the NBP estimations, both the non-weighted optimization (NW, *Eq. 5, 6*) and the distance weighted optimization (DW, *Eq. 9, 10*) were performed. The optimization of Group Contributions for each estimation task without weighting (NW) is denoted by JR-NW and MG-NW, with Distance Weighting (DW) by JR-DW and MG-DW.

The group matrix $[n]$ was modified for the calculations of the Group Contribution coefficients of MG-NW & MG-DW methods. The column corresponding to group ($>C<$) was removed to avoid matrix singularity, *i.e.* its GC value was set to 0. The singularity appears due to restriction of the database of the MG method to alkanes and alcohols, since the amount of end groups $-CH_3$ and $-OH$ is a linear function of the number of $>C<$ and $-CH<$ groups (*i.e.* $n_{CH_3} + n_{OH} = 2 + n_{CH} + 2n_C$). This modification of matrix $[n]$ does not influence the NBPs estimations and it is not needed when cyclic compounds are included in the database.

3 RESULTS AND DISCUSSION

3.1 NBP Estimations

The performances of the NW and the DW technique based on the JR and the MG methods were compared with respect to NBP estimation. The Absolute Error (AE) was calculated as the absolute difference between the predicted NBP and the experimental value. The Absolute Error frequency distribution for JR, JR-NW and JR-DW is shown in Figure 1. The number of estimations with AE between given values (for example between 0 and 1) is shown between those values on abscissa. The AE of both the JR and the JR-NW methods had a rather similar frequency distribution as can be seen in Figure 1. About 30 NBP estimations of the JR method had high deviations (>40 K). The NW method behaved similarly but the error distribution was slightly more even. This was expected since the optimization procedure in JR-NW minimizes the deviations based on the standard deviation residual function; whereas JR is based on Absolute Average Errors (AAE) (Joback and Reid, 1987). The use of weights in JR (JR-DW) reduced the AAE by 6.3 K compared to JR-NW. Moreover, the number of extremely high deviations was reduced considerably, only 8 cases were found with deviation above 40 K (see *Fig. 1*). Figure 1b shows the exponential approximation of the AE frequencies *versus* AE intervals (drawn as Microsoft Excel trendline utilizing least square optimization of the coefficients for the exponential line based on all experimental points). The slope of the JR-DW frequency trend line is much steeper, *i.e.* JR-DW predicts more accurately. The AAE of JR-DW was 8.9 K, whereas for JR it

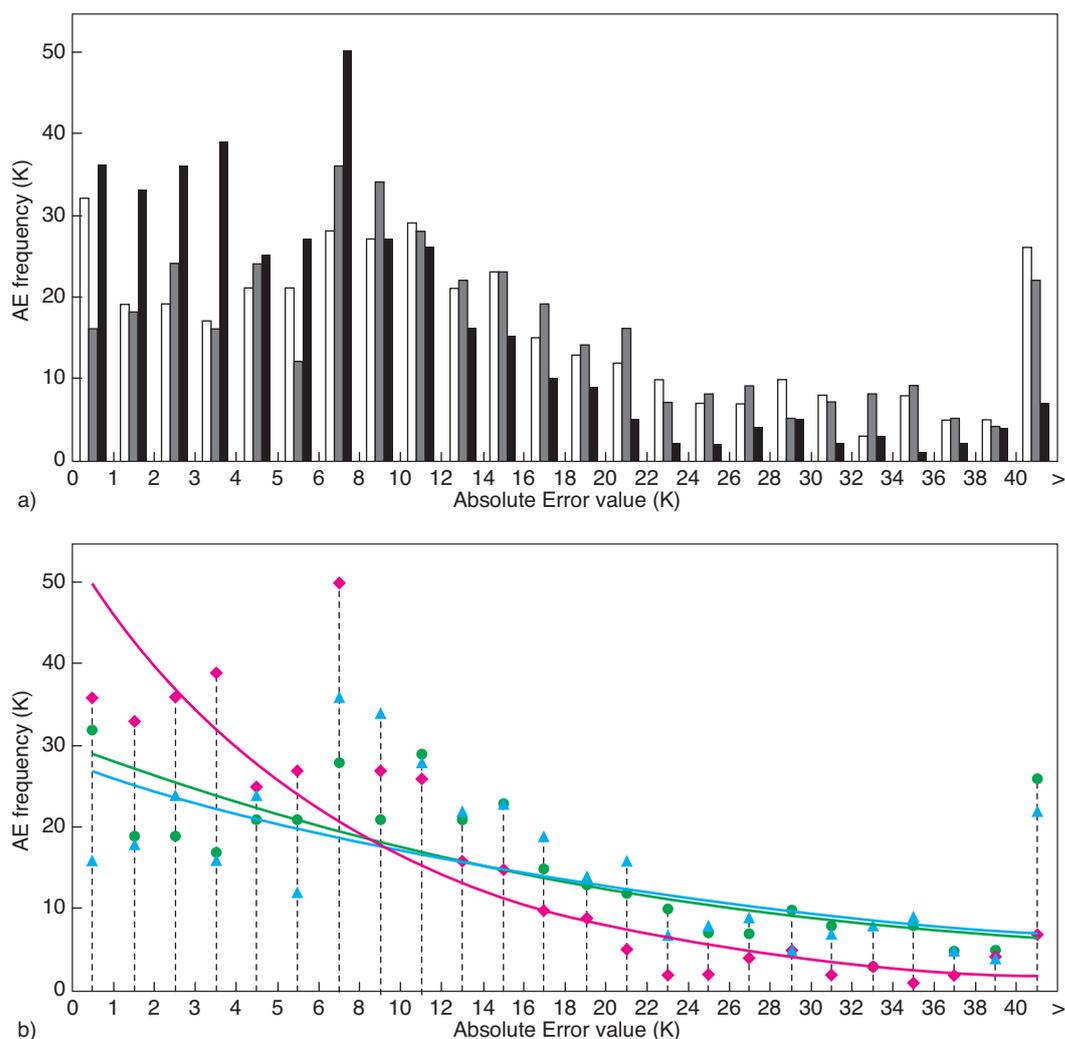


Figure 1

a) Absolute Error frequency (K) of measured (Poling *et al.*, 2001) and estimated NBP calculated with the JR method (open bars □), with the JR-NW method (with Eq. 5 gray bars ■) and with the JR-DW method (with Eq. 9, black bars ■) versus interval of the absolute error; b) Exponential trend lines for the frequency data, where green circles (●) are JR AE frequencies and its trend line (green line); blue triangles (▲) are JR-NW AE frequencies and its trend line (blue line); red diamonds (◆) are JR-DW AE frequencies and its trend line (red line).

was 15.7 K and for JR-NW 15.2 K. The direct comparison between JR-DW and JR is improper, because of the difference in the databases, in the minimization techniques and in the procedures for property estimation. On the other hand, the results of JR-NW and JR are expected to be similar. Therefore, the comparison between JR-NW and JR-DW properly reveals the ability of distance weighting of improving the used GCM.

The poor performance of JR method for NBP estimation is well known (Poling *et al.*, 2001); other GCMs could improve the NBP estimation by exploiting different functional form of NBP dependency on groups (Cordes and Rarey, 2002; Marrero and Gani, 2001) and/or by explicit use of the molecular weight in the NBP estimation (Marrero-

Morejon and Pardillo-Fontdevila, 1999; Nannoolal *et al.*, 2004). These methods can be used as a basis of the DW technique as well. But even for the JR method, increasing weight of similar compounds in the GC optimization allows taking into account implicitly many phenomena that can play a role in vaporization process, such as molecule volume, surface tension, molecular flexibility, etc. At the same time the JR-DW method keeps very simple form of the JR method.

The frequencies of the absolute errors of MG, MG-NW and MG-DW are shown in Figure 2. A reduction of the extremely high deviations is also observed with the MG-DW method applied to alkanes and alcohols. The MG-DW NBP estimations were on average about 4 K closer to the experimental NBP values than the MG-NW method estimations.

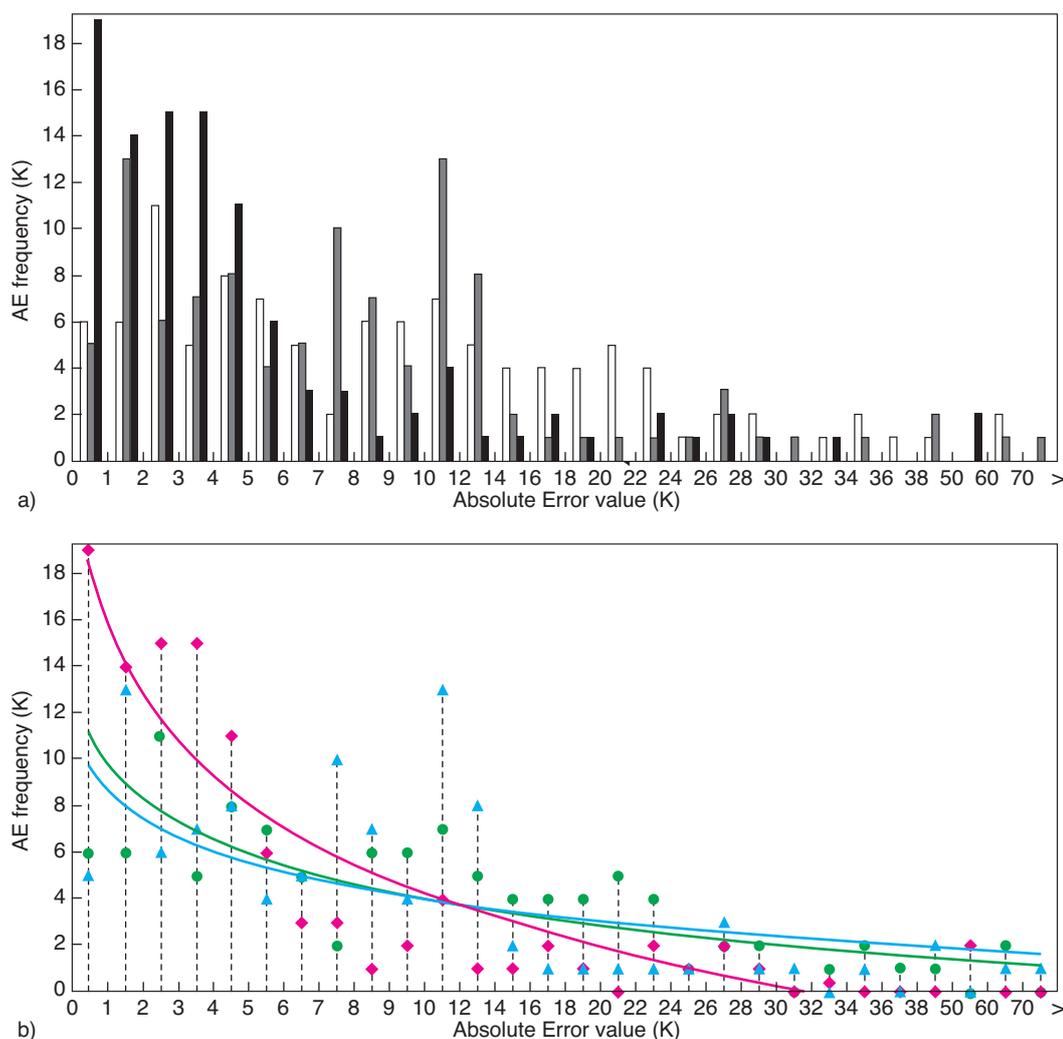


Figure 2

a) Absolute Error frequency (K) of measured (Poling *et al.*, 2001) and estimated NBP calculated with the MG method (open bars \square), with the MG-NW method (with Eq. 5, 6, gray bars \blacksquare) and with the MG-DW method (with Eq. 9, 10, black bars \blacksquare); b) Exponential trend lines for the frequency data, where green circles (\bullet) are JR AE frequencies and its trend line (green line); blue triangles (\blacktriangle) are JR-NW AE frequencies and its trend line (blue line); red diamonds (\blacklozenge) are JR-DW AE frequencies and its trend line (red line).

For both GCMs, the greater improvements are obtained for small and large molecules (C_1 - C_3 , C_{21} - C_{24}), with a gain in accuracy of 12.6 K and 12 K for JR-DW and MG-DW respectively (see also Fig. 3, 4). In Figure 3, the NBP estimations with NWs and DWs are plotted against the experimental NBPs. The NW slightly improves the NBP estimations of both JR and MG by reducing the deviations of the high boiling compounds. The improvement of the estimation with the DW methods can be clearly seen in Figure 4, where errors of the NBP estimations are plotted against the experimental NBP for both the JRs and MG methods. The deviation trend of JR-NW and MG-NW is similar to original GCMs, *i.e.* larger deviations for low and high boiling compounds. The DW

methods clearly reduce all deviations, without compromising the quality of the estimations in the middle boiling region (see Fig. 4). The gain in the estimation accuracy with the DW methods is clearly visible also in Figure 5. We observed a steady decrease of the deviation fraction and a sharper gain in accuracy in the high deviation range for both JR-DW (Fig. 5a) and MG-DW (Fig. 5b).

The quality of the performances of the proposed methods depends on the available experimental data for the compounds containing each considered structural group. An overview of the databases used in this work is given in Table 2. The Average Absolute Errors (AAE) per class of compounds are included in the table.

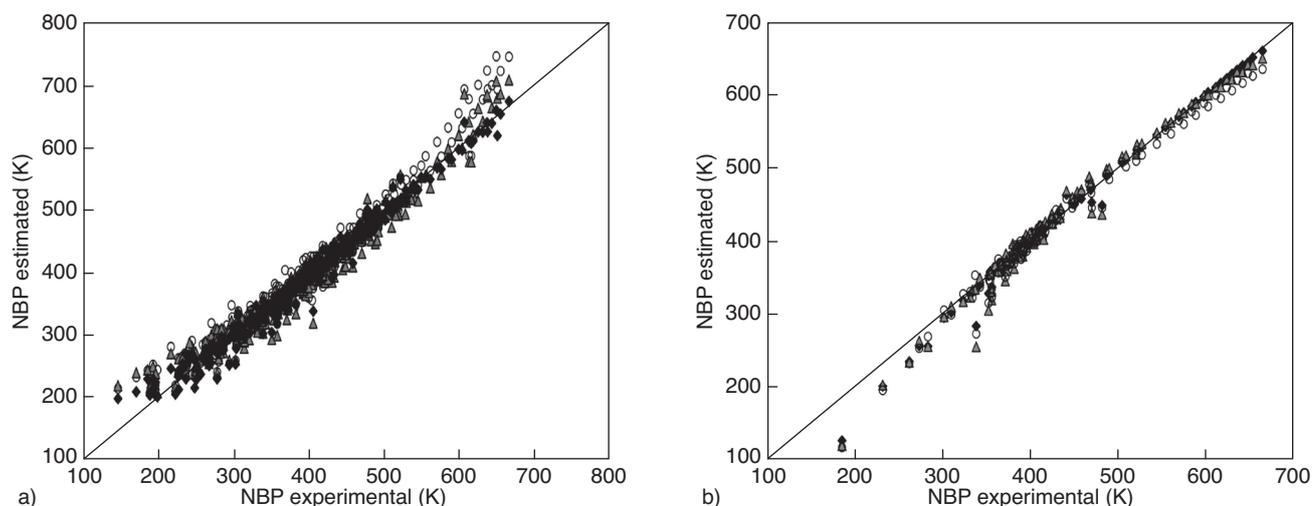


Figure 3

NBP estimation calculated with different methods *versus* the experimental NBP. a) The JR (○), the JR-NW method (▲) and the JR-DW method (◆); b) The MG (○), the MG-NW method (▲) and the MG-DW method (◆).

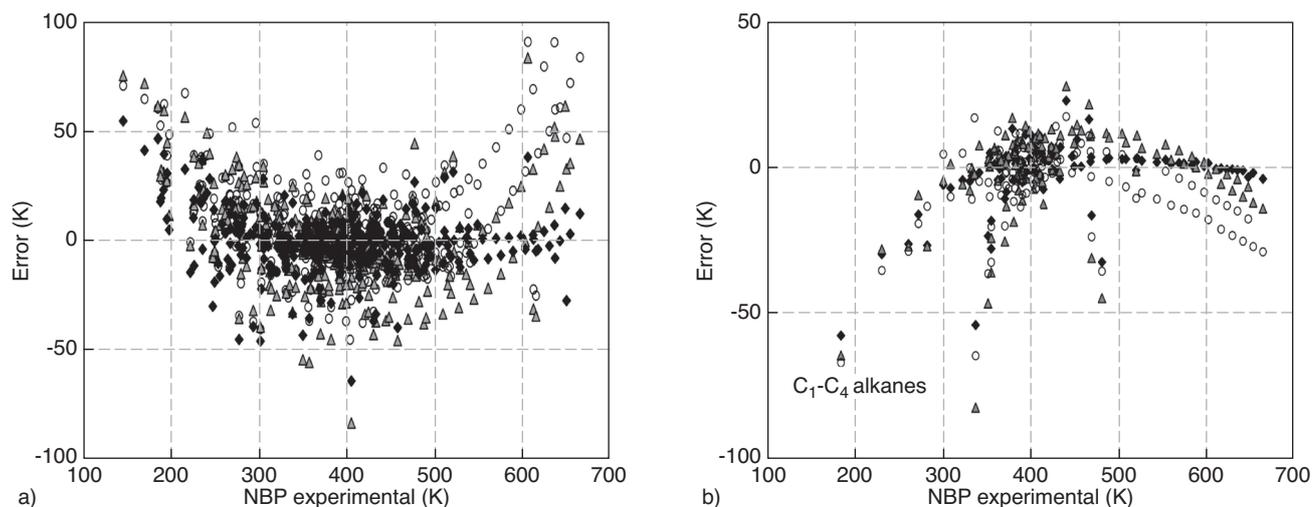


Figure 4

NBP estimation error calculated with different methods *versus* the experimental NBP. a) The JR (○), the JR-NW method (▲) and the JR-DW method (◆); b) The MG (○), the MG-NW method (▲) and the MG-DW method (◆).

Clearly, the JR-DW method performed considerably better than the JR-NW method with respect to the NBP estimations. The JR-DW performs either better or comparably to the original JR method. With all methods, very high AAE was observed for some classes of compounds, *i.e.* alkynes and Br containing compounds. The low accuracy for these classes derives from the characteristics of our database: only two compounds containing Br without any other halogen were included in the database, and for alkynes, the triple bond

groups were independent only in the case of ethyne (CH# group) and of dimethylacetylene (–C# group).

The AAEs of the MG-NW and the MG-DW methods are presented in Table 2. MG-DW performs better than both the MG-NW and the original MG methods, thus using distance weighting improves NBP estimations regardless of the database size. Additionally, application of DW approach decreased difference in the estimations of two tested GCMs, *i.e.* AE for JR-DW and MG-DW estimations of alkane NBPs were

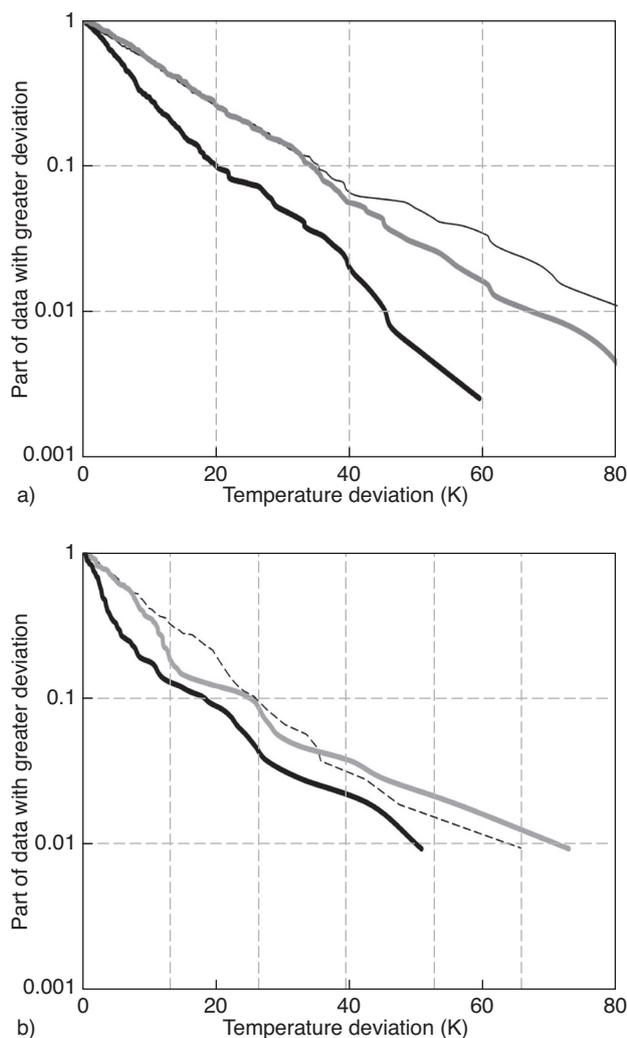


Figure 5

Fraction of the data with deviations larger than a given value.
 a) Plot is based on 386 compounds database; the JR method (—), the JR-NW method (—) and the JR-DW method (—);
 b) Plot is based on alkanes and alcohol database (107 compounds); the MG method (---), the MG-NW method (—) and the MG-DW method (—).

7.4 and 5 K correspondingly, whereas the estimation with the original methods gave larger difference (the AAE were 14.6 and 9.9 for the JR and the MG methods correspondingly).

In general, for all classes of compounds the DW method was superior to the NW method.

3.2 Changes in the Group Contribution Values within Estimation of NBP of Homological Series of 1-Alcohols

The proposed weighting strategy assumes that there is no standard contribution for each group, *i.e.* the Group

TABLE 2

The average absolute error of NBP estimation of different compound classes in the selected database

Group	Sub-group	$N^{(1)}$	GC method		
			JR	JR-NW	JR-DW
Alkanes		67	14.6	15.2	7.4
Alkenes		19	18.3	24.4	9.5
Alkynes		5	9.1	19.9	11.5
Halogen containing compounds	F	34	24.0	19.7	14.9
	Cl	11	16.3	25.9	14.4
	Br	2	13.4	51.7	37.9
	Multihalogen & polyfunctional	26	28.5	18.6	12.2
	Overall	74	24.2	21.1	14.5
Organic acids		12	10.6	7.3	5.5
Alcohols	Primary	20	34.1	22.5	6.8
	Overall	44	24.1	16.4	8.2
Esters	Formates	6	9.8	5.4	2.2
	Acetates	7	4.4	5.2	4.5
	Butanoates & propanoates	17	6.9	7.4	4.4
	Other esters	3	10.7	9.8	4.3
	Overall	33	7.2	6.8	4.0
Cyclic hydrocarbons		21	12.2	16.2	9.0
Aromatics	Benzenes/toluenes /napthalenes	34	11.8	13.7	7.1
	Aromatic amines	13	9.1	9.6	8.6
	Aromatic alcohols	14	7.3	8.8	7.9
	Others	3	11.0	8.9	5.8
	Overall	64	10.6	12.3	8.1
Ethers		7	10.7	13.9	9.4
Ketones		12	6.2	7.8	5.5
Epoxides		6	12.9	17.0	11.1
Nitrogen containing		16	14.5	14.6	9.5
Sulfur containing		6	4.8	10.2	4.5
Overall		386	15.7	15.2	8.9
			MG	MG-NW	MG-DW
Alkanes	<i>n</i> -alkanes	23	15.5	11.3	5.7
	Overall	67	9.9	8.1	5.0
Alcohols	Primary	20	12.2	13.6	2.4
	Overall	40	11.0	14.5	8.7
Overall		107	10.32	10.48	6.42

(1) Number of components.

Contributions (GC) are not stored. They are newly optimized for each estimation task based on the available database. In this work, we tested the change in value of the GCs for 1-alcohols in NBP estimations. Figure 6 shows that the Non-Weighted method (NW) does not change considerably

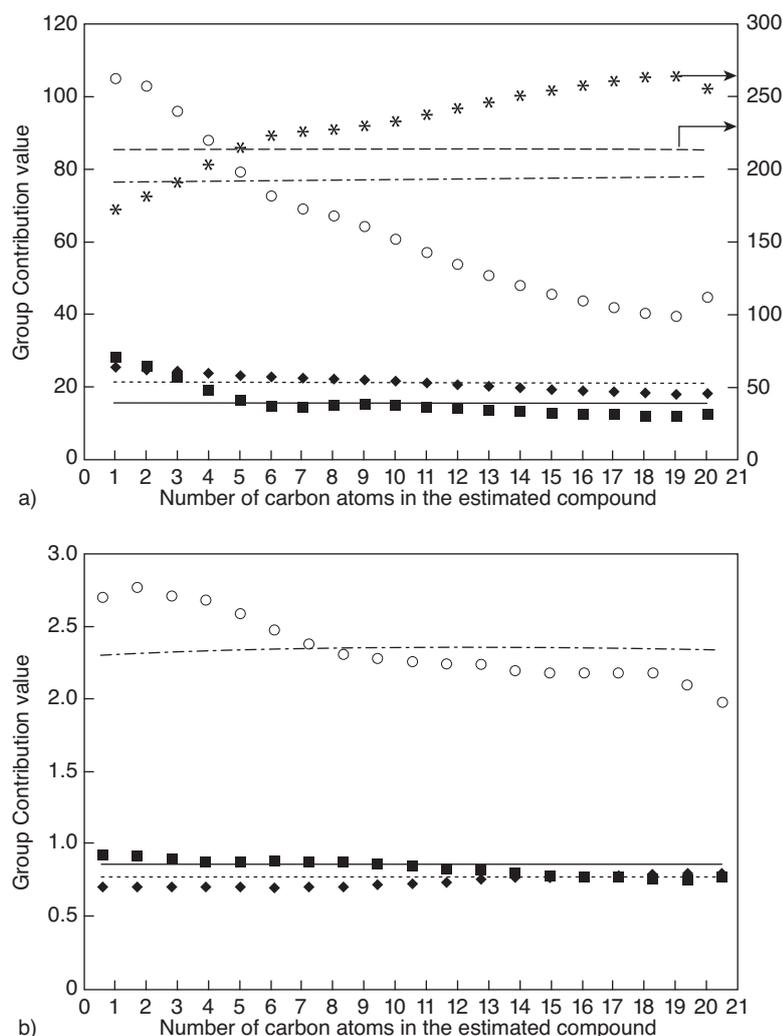


Figure 6

Shift in the optimized 1-st order group contribution values when estimating 1-alcohols NBP by the JR and the MG methods. 1-st order groups; values optimized within NW method $-\text{CH}_3$ (—), $-\text{CH}_2$ (- - -), $-\text{OH}$ (- - -); a_0 value (- - -); values optimized within DW method: $-\text{CH}_3$ (■); $-\text{CH}_2$ (◆); $>\text{CH}$ (▲), $-\text{OH}$ (○), a_0 value (*). a) Based on the JR method; b) based on the MG method.

the GC values. With the DW method, the GC values change depending on the estimated compound due to the differences in weighting during the optimization. For $-\text{CH}_3$ and $-\text{CH}_2$ groups the change is very small, while for the $-\text{OH}$ group it is much more prominent (Fig. 6). The GC of the $-\text{OH}$ group becomes smaller in value for larger alcohols. When the group independent constant is optimized, *i.e.* in the JR-NW and in the JR-DW, the value of this constant increases with increasing carbon number of the estimated alcohol. The other group contributions slightly decrease in value (Fig. 6a).

Trends of change for second order group contribution values optimized in the MG-NW and the MG-DW methods are more complex. The variation of the values are influenced by weighting and by the changes in contributions derived from the optimization of the first order groups. Generally, the

second order group contribution values increase slightly with increasing carbon number of the estimated alcohol.

3.3 Estimation of Other Physical Properties

Besides Normal Boiling Point, the proposed technique was also tested on estimating other thermodynamic properties of non-electrolyte compounds employing the JR GCM. Standard enthalpy and Gibbs energy of formation, melting temperature, enthalpy of fusion and vaporization as well as critical pressure and volume were estimated. A list of properties with the used equations is given in Table 3. The critical temperature was not calculated since it cannot be linearized in its JR representation. One can see that the accuracy of the

TABLE 3

Physical properties estimations with NW and DW methods based on the JR GCM

Property	Linearized equation	N	AAE		
			JR	JR-NW	JR-DW
T_b (K) Normal Boiling Point	$T_b = \sum_{i=0}^{N_G} a_i n_i, n_0 = 1$	386	15.7	15.2	8.9
P_c (bar) critical pressure	$(\sqrt{P})^{-1} = \sum_{i=0}^{N_G} a_{P_c}^i n_i^{**}, **$	294	2.0 (0.4) [†]	2.1 (0.5) [†]	1.6 (0.3) [†]
V_c (cm ³ /mol) critical volume	$V_c = \sum_{i=0}^{N_G} a_{V_c}^i n_i^{**}$	266	17.9 (2.0) [†]	18.4 (2.4) [†]	11.4 (1.5) [†]
T_m (K) melting temperature	$T_m = \sum_{i=0}^{N_G} a_{T_m}^i n_i^{**}$	309	27.5	28.9	26.0
H_{fus} (kJ/mol) fusion enthalpy	$H_{fus} = \sum_{i=1}^{N_G} a_{H_{fus}}^i n_i^{**}$	132	2.7	3.1	2.9
$H_{f,298}$ (kJ/mol) standard enthalpy of formation	$H_f = \sum_{i=0}^{N_G} a_{H_f}^i n_i^{**}$	314	10.7	12.6	8.6
$G_{f,298}$ (kJ/mol) standard Gibbs energy of formation	$G_f = \sum_{i=0}^{N_G} a_{G_f}^i n_i^{**}$	266	11.4	10.7	7.8
H_{vap} (kJ/mol) enthalpy of vaporization	$H_{vap} = \sum_{i=0}^{N_G} a_{H_{vap}}^i n_i^{**}$	169	1.5	1.8	1.3

$$* (\sqrt{P})^{-1} = 0.113 + 0.0032n_A - \sum_{i=1}^{N_G} a_{P_c}^i n_i = 0.113 + \sum_{i=1}^{N_G} (0.0032 - a_{P_c}^i) n_i = \sum_{i=0}^{N_G} a_{P_c}^i n_i,$$

n_A : total number of non-hydrogen atoms.

Constrained quadratic programming problems minimization was used to find only coefficients that produce non-negative pressure.

** $n_0 = 1$, a_0 : group independent property constant.

† Absolute Average Error calculated per heavy (not hydrogen) atoms in the molecules.

methods reported in Table 3 is relatively low. This is due to both the original JR method predictive ability and to practical difficulties in experimental determination of some of the properties (Katritzky *et al.*, 2010), such as a fusion enthalpy.

The database by Poling *et al.* (2001) also includes values of certain properties that were calculated with an estimation method. These values were not used in this work, thus there are different amount of available points (N) for GC optimization of different physical properties.

The JR-DW method was always superior to the JR-NW method in predicting any of the considered physical properties.

The present study was carried out with a database that is relatively small compared to those used in typical process flowsheet simulators or in other advanced thermodynamic estimation packages. In larger databases, there are inherently more components with closely similar chemical character; thus, the DW technique can even further improve predictions if applied to a larger database.

CONCLUSIONS

A novel approach in the use of GCMs was developed in this work. This methodology takes into account the structural similarities among compounds in the optimization of the group contribution coefficients by means of weighting factors. The Distance Weighting (DW) method relies on an available database of experimental physical properties. The characteristics of the database are very important for the performance of the DW method. In fact, the accuracy of property estimations for a compound is strongly dependent on the amount of compounds with the same structural groups included in the database. The method increases the impact of specific groups on the estimations; thus using structural groups that reveal important characteristics of a molecule is essential to the success of the method.

The DW method could be applied to any GCM and it improves the GCM to which it is applied. In this work, the DW method was applied to the Joback and Reid and to the

Marrero and Gani GCMs (Joback and Reid, 1987; Marrero and Gani, 2001). The JR method was used to estimate NBP and other physical properties of compounds for which experimental data were found in Poling *et al.* (2001). The method of Marrero and Gani was used for NBP estimations of alkanes and alcohols from the same database. The DW method reduced the Absolute Average Error (AAE) in NBP of by 6.3 K for the JR method and of by 4 K for the MG method.

The advantages of this method can be exploited in several ways. The DW method can be used for the fast optimization of properties or model parameters of a compound based on a relatively small database, in which a sufficient number of similar-in-structure compounds are present. The database might even be restricted to specific classes of chemicals. The DW method can also be applied to advanced GCMs using a large database. In this case, the benefits from detailed group definition and from the use of second and third order groups are integrated.

ACKNOWLEDGMENTS

This work was supported by National Technology Agency of Finland (TEKES). The authors are grateful to Dr. Claudia Dell’Era for valuable discussion and help in preparation of the paper.

REFERENCES

- Constantinou L., Gani R. (1994) New group contribution method for estimating properties of pure compounds, *AICHE J.* **40**, 10, 1697-1709.
- Cordes W., Rarey J. (2002) A new method for the estimation of the normal boiling point of non-electrolyte organic compounds, *Fluid Phase Equilib.* **201**, 2, 409-433.
- DIPPR Project 801 (2005/2008/2009) Design institute for physical property data/AIChE.
- Dudani S.A. (1976) The distance-weighted k-nearest-neighbor rule, *IEEE Trans. Syst. Man Cybern.* **SMC-6**, 4, 325-327.
- Gani R., Harper P.M., Hostrup M. (2005) Automatic creation of missing groups through connectivity index for pure-component property prediction, *Ind. Eng. Chem. Res.* **44**, 18, 7262-7269.
- Gmehling J. (2009) Present status and potential of group contribution methods for process development, *J. Chem. Thermodyn.* **41**, 6, 731-747.
- González H.E., Abildskov J., Gani R., Rousseaux P., Bert B.L. (2007) A method for prediction of UNIFAC group interaction parameters, *AICHE J.* **53**, 6, 1620-1632.
- Joback K.G., Reid R.C. (1987) Estimation of pure-component properties from group contributions, *Chem. Eng. Commun.* **57**, 1-6, 233-243.
- Kang J.W., Diky V., Chirico R. D., Magee J. W., Muzny C.D., Abdulagatov I., Kazakov A.F., Frenkel M. (2011) A new method for evaluation of UNIFAC interaction parameters, *Fluid Phase Equilib.* **309**, 1, 68-75.

Katritzky A.R., Kuanar M., Slavov S., Hall C.D., Karelson M., Kahn I., Dobchev D.A. (2010) Quantitative correlation of physical and chemical properties with chemical structure: Utility for prediction, *Chem. Rev.* **110**, 10, 5714-5789.

Kolská Z., Petrus P. (2010) Tool for group contribution methods – computational fragmentation, *Collect. Czech. Chem. Commun.* **75**, 4, 393-404.

Marrero J., Gani R. (2001) Group-contribution based estimation of pure component properties, *Fluid Phase Equilib.* **183-184**, 183-208.

Marrero-Morejon J., Pardillo-Fontdevila E., Fernandez-Benitez S. (1999) Estimation of pure compound properties using group-interaction contributions, *AICHE J.* **45**, 3, 615-621.

Marrero-Morejón J., Pardillo-Fontdevila E. (2000) Estimation of liquid viscosity at ambient temperature of pure organic compounds by using group-interaction contributions, *Chem. Eng. J.* **79**, 1, 69-72.

Marrero-Morejon J., Pardillo-Fontdevila E. (1999) Estimation of hydrocarbon properties from group-interaction contributions, *Chem. Eng. Commun.* **176**, 161-173.

Mustaffa A.A., Kontogeorgis G.M., Gani R. (2011) Analysis and application of GCPlus models for property prediction of organic chemical systems, *Fluid Phase Equilib.* **302**, 1-2, 274-283.

Nannoolal Y., Rarey J., Ramjugernath D., Cordes W. (2004) Estimation of pure component properties: Part 1. Estimation of the normal boiling point of non-electrolyte organic compounds via group contributions and group interactions, *Fluid Phase Equilib.* **226**, 45-63.

Peters F.T., Laube F.S., Sadowski G. (2012) Development of a group contribution method for polymers within the PC-SAFT model, *Fluid Phase Equilib.* **324**, 0, 70-79.

Poling B.E., Prausnitz J.M., O’Connell J.P. (2001) *Properties of gases and liquids*, Poling B.E., Prausnitz J.M., O’Connell J.P. (eds), 5th edition, McGraw-Hill.

Raymond J.W., Rogers T.N. (1999) Molecular structure disassembly program (MOSDAP): A chemical information model to automate structure-based physical property estimation, *J. Chem. Inf. Comput. Sci.* **39**, 3, 463-474.

Rowley J., Wilding W., Oscarson J., Rowley R. (2007) Rapid evaluation of prediction methods with DIPPR’s automated property prediction package, *Int. J. Thermophys.* **28**, 3, 824-834.

Soria T.M., Andreatta A.E., Pereda S., Bottini S.B. (2011) Thermodynamic modeling of phase equilibria in biorefineries, *Fluid Phase Equilib.* **302**, 1-2, 1-9.

Tochigi K., Yoshida K., Kurihara K., Ochi K., Murata J., Urata S., Otake K. (2002) Determination of ASOG parameters for selecting azeotropic mixtures containing hydrofluoroethers, *Fluid Phase Equilib.* **194-197**, 0, 653-662.

Vijande J., Piñeiro M.M., Legido J.L., Bessièrès D. (2010) Group-contribution method for the molecular parameters of the pc-saft equation of state taking into account the proximity effect. Application to nonassociated compounds, *Ind. Eng. Chem. Res.* **49**, 19, 9394-9406.

Wakeham W.A., Cholakov G.St., Stateva R.P. (2002) Liquid density and critical properties of hydrocarbons estimated from molecular structure, *J. Chem. Eng. Data* **47**, 3, 559-570.

Wang Q., Ma P., Wang C., Xia S. (2009) Position group contribution method for predicting the normal boiling point of organic compounds, *Chinese J. Chem. Eng.* **17**, 2, 254-258.

Final manuscript received in July 2012
Published online in March 2013

Copyright © 2013 IFP Energies nouvelles

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 Energies nouvelles 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 Information Mission, IFP Energies nouvelles, fax. +33 1 47 52 70 96, or revueogst@ifpen.fr.

APPENDIX

Let's consider estimation of the NBP for 2-methyl-2-butanol by methods described in our article. Its experimental NBP value is 375.15 K.

The JR, JR-NW and JR-DW methods

2-methyl-2-butanol consists of the following JR groups (n_i): $3 \times -CH_3$, $1 \times -CH_2$, $1 \times >C<$, $1 \times -OH$ (aliphatic).

Equation (2) is used for the NBP calculation in the JR method, *i.e.*:
$$NBP_{calc} = a_0 + \sum_{i=1}^{N_G} a_i n_i$$

Values of the original JR group contributions are given in the table together with the estimated NBP.

At first Equation (5) was applied to the whole database except 2-methyl-2-butanol (386 compounds from Poling book). As a result 40 values of the Group Contributions for JR groups were obtained. Five of those values were used for estimation of 2-methyl-2-butanol NBP (see *Tab. A*).

Then, weighted GC optimization was made in accordance with Equation (9) with the same database. Five of forty optimized GC values were used for estimation of the NBP by the JR-DW. The GC and estimated NBP are given in Table A.

The MG, MG-NW and MG-DW methods

2-methyl-2-butanol consists of the following first order MG groups (n_i): $3 \times -CH_3$, $1 \times -CH_2$, $1 \times >C<$, $1 \times -OH$ (aliphatic), and the second order group (COH group).

The NBP of 2-ethyl-2-butanol is calculated in accordance with following equation:
$$NBP_{calc} = T_{b0} \cdot LN \left(\sum_{i=1}^{N_G} a_i n_i \right)$$
 where constant T_{b0} was used as it is given in the original MG method, *i.e.* 222.543.

The original MG group contributions and estimated NBP is given in Table B.

For the MG-NW method, Equations (5) and (6) were used for optimization of GC values with the database of alkanes and alcohols (see *Sect. 1.1 and 2.1*). Exponential function of relative boiling temperature ($\exp(T_b/T_{b0})$) was used as the estimated property for linearization. Eleven group contributions were optimized and four of them used for estimation of 2-methyl-2-butanol NBP. It is worth to remind that GC value for $>C<$ group was not optimized as it is described in Section 2.2.

Equations (9) and (10) were used for the MG-DW method. The resulting optimized GC and estimated with them NBP is shown in Table B.

In both shown examples utilization of the distance weighting improves the NBP estimation.

TABLE A
Estimation of NBP of 2-methyl-2-butanol with the JR, JR-NW and JR-DW methods

	a_0	$-CH_3$	$>CH_2$	$>C<$	$-OH$ (a)	$\sum_{i=1}^{N_G} a_i n_i$	NBP (K)	AE (K)
JR groups, n_i		3	1	1	1			
JR, a_i	198.2	23.58	22.88	18.25	92.88	204.8	403.0	27.9
JR-NW, a_i	213.9896	15.585	20.9912	35.7654	77.4111	180.9	394.9	19.8
JR-DW*, a_i	186.3641	24.8815	24.9821	13.5648	88.8115	202.0	388.4	13.2

* Note that the provided JR-DW GC values can be used only for estimation of 2-methyl-2-butanol NBP, because they were optimized with specific to 2-methyl-2-butanol weights as described in Section 1.2.

TABLE B
Estimation of NBP of 2-methyl-2-butanol with the MG, MG-NW and MG-DW methods

	$-CH_3$	$>CH_2$	$>C<$	$-OH$ (a)	COH	$\sum_{i=1}^{N_G} a_i n_i$	NBP (K)	AE (K)
MG groups, n_i	3	1	1	1	1			
MG, a_i	0.8491	0.7141	-0.0671	2.567	-0.5325	5.2288	368.1	7.0
MG-NW, a_i	0.8635	0.772	-	2.3503	-0.2993	5.4959	379.2	4.1
MG-DW*, a_i	0.8817	0.712	-	2.4725	-0.4161	5.4688	378.1	3.0

* Note that the provided MG-DW GC values can be used only for estimation of 2-methyl-2-butanol NBP, because they were optimized with specific to 2-methyl-2-butanol weights as described in Section 1.2.