Journal of Chromatography A, 1425 (2015) 249–257

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Robust estimation of thermodynamic parameters (ΔH, ΔS and ΔC_p) for prediction of retention time in gas chromatography – Part I (Theoretical)

Carlos Alberto Claumann[a], André Wüst Zibetti[c,*], Ariovaldo Bolzan[a], Ricardo A.F. Machado[a], Leonel Teixeira Pinto[b]

[a] Laboratório de Controle de Processos, Departamento de Engenharia Química e de Alimentos, Centro Tecnológico, Universidade Federal de Santa Catarina (UFSC), P.O. Box: 476, Florianópolis, SC 88010-970, Brazil
[b] Laboratório de neuroengenharia computacional, Departamento de Engenharia Química e de Alimentos, Centro Tecnológico, Universidade Federal de Santa Catarina (UFSC), P.O. Box: 476, Florianópolis, SC 88010-970, Brazil
[c] Departamento de Informática e Estatística (INE), Centro Tecnológico, Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC 88010-970, Brazil

ARTICLE INFO

Article history:
Received 11 August 2015
Received in revised form 19 October 2015
Accepted 20 October 2015
Available online 6 November 2015

Keywords:
Parameter estimation
Initialization method
Gas chromatography
Thermodynamic properties
Retention time
Optimization

ABSTRACT

An approach that is commonly used for calculating the retention time of a compound in GC departs from the thermodynamic properties ΔH, ΔS and ΔC_p of phase change (from mobile to stationary). Such properties can be estimated by using experimental retention time data, which results in a non-linear regression problem for non-isothermal temperature programs. As shown in this work, the surface of the objective function (approximation error criterion) on the basis of thermodynamic parameters can be divided into three clearly defined regions, and solely in one of them there is a possibility for the global optimum to be found. The main contribution of this study was the development of an algorithm that distinguishes the different regions of the error surface and its use in the robust initialization of the estimation of parameters ΔH, ΔS and ΔC_p.

1. Introduction

Different approaches have been proposed for predicting retention time of compounds in a capillary column in gas chromatography, and the main one is based on thermodynamic properties (ΔH, ΔS and ΔC_p) of analytes [1–6]. The estimation of these parameters in gas chromatography can be obtained by isothermal runs, and such estimate presents analytical solutions [7,8]. Another way to estimate these parameters is through the use of different temperature ramps, which can considerably reduce the number of experiments [9]. In this case there is no analytical solution, and the problem must be solved with the aid of numerical methods, using non-linear optimization techniques in order to estimate the parameters from a number of temperature programs [10,11]. So as to solve the non-linear optimization problem, different routines can be used, and for each class of problem a specific method may be preferable. Among some of these routines, those based on differential calculus, and those based on computational intelligence can be mentioned [12–16]. These procedures may present different performances in terms of convergence speed, and, proportionally, the number of assessments of the objective function. The methods based on differential calculus show strong dependence on initial conditions (initial guess), or the starting point of the algorithm. This dependence on initial condition means that the optimal solution is not guaranteed most of the time, which requires several evaluations with different initial conditions from the modeler. Nonetheless, such methods may be able to converge quickly, depending on how the problem in question was proposed, on the type of objective function, and a good initial guess. Thus, the choice of a starting point determines how quickly the algorithm converges to a solution. However, some methods based on artificial intelligence are more robust in terms of initial conditions, but can be, in some cases, highly demanding, computationally [17,18]. Similarly to the calculus-based methods, there is a strong need for checking convergence. In both cases, repeated assessments of the process are common practice among modelers in the search for the global optimum.

* Corresponding author.
E-mail addresses: carlos.claumann@prosgrad.ufsc.br (C.A. Claumann), azibetti@gmail.com (A. Wüst Zibetti), ariovaldo.bolzan@ufsc.br (A. Bolzan), ricardo.machado@ufsc.br (R.A.F. Machado), leonel.t.pinto@ufsc.br (L.T. Pinto).

http://dx.doi.org/10.1016/j.chroma.2015.10.072
0021-9673/

In this light, one way to increase performance for parameter estimation is to improve the initialization of the calculations-based methods. Thus, robustness and convergence to the global optimum with a small number of objective function evaluations would be obtained. The robust initialization (proposed in this paper) consists of finding out which region is closer to the global optimum. In this case, it is necessary to know the particularities of the surface of the objective function with respect to the parameters that need adjustment.

This work demonstrates that the approximation error of the objective function can be split into different regions, according to variations of the values of thermodynamic parameters. There are basically three main regions, and only in one of them is there a possibility for the global optimum to be found.

As will be shown, an error criterion (for example, the sum of the square error) can assume very high values, or tend toward a constant value, depending on the considered area in terms of thermodynamic parameters. Such areas indicate large and small parametric sensitivity, respectively.

The main contribution of the work we have put forth is the mathematical proof that the error criterion can be divided into three specific regions, only one of which shows a possibility for the global optimum to be found. Another important contribution was the development of an algorithm that discriminates between the different regions of the error criterion surface, as well as the development of a strategy for initializing parameter estimation, thus generating a point exactly in the region where the global optimum can be found. This helps establish the possibility for robust initialization for the estimation of parameters ΔH, ΔS and ΔC_p. Additionally, the regions found can be used for defining restrictions related to the problem of parameter estimation. Moreover, both the mathematical proof and the definition of these regions are valid for any analyte regardless of the stationary phase.

So as to observe the different regions of the approximation error criteria, the sum of the square error as objective function was used; however, it should be noted that other criteria could be used, such as the maximum error in absolute value, or the sum of the errors in absolute value.

2. Prediction of time retention in CG model

In GC, the solvation of an analyte placed within the carrier gas inside a capillary column with a specific stationary phase is expressed by the relationship between $\Delta G(T)$ and the distribution coefficient $K(T)$. The dependency the temperature will bear to ΔG can be established by the basic thermodynamic relationship in terms of $\Delta H(T_0)$, $\Delta S(T_0)$ and ΔC_p. The $\Delta H(T_0)$, $\Delta S(T_0)$ and ΔC_p represent the changes in enthalpy and entropy associated with the transfer of the solute from the mobile to the stationary phase, at a given temperature T_0; T_0 is an arbitrary reference temperature, and the one chosen for this work 50 °C (typically assumed to be equal to T used for time $t = 0$), and ΔC_p is the change in its isobaric heat capacity for the transfer. The thermodynamic relationship is expressed in $K(T)$ as,

$$\Delta H(T(t)) = \Delta H(T_0) + \Delta C_p(T(t) - T_0) \tag{1}$$

$$\Delta S(T(t)) = \Delta S(T_0) + \Delta C_p(\ln T(t) - \ln T_0) \tag{2}$$

$$\ln K(T(t)) = -\frac{\Delta H(T(t))}{RT(t)} + \frac{\Delta S(T(t))}{R} \tag{3}$$

where R is the gas constant.

The distribution coefficient K is related to retention factor k according to Eq. (4). In that case, Eq. (3) can be expressed directly in terms of k as shown in Eq. (5).

$$k(T(t)) = \frac{K(T(t))}{\beta} \tag{4}$$

where β is the phase ratio of the column.

$$\ln\ k(T(t)) = -\frac{\Delta H}{RT(t)} + \frac{\Delta S}{R} - \ln\beta \tag{5}$$

If the dead time t_M can be described as a function only of time (but not the position within the column), it is possible to apply separation of variables to the differential equation describing the motion of an analyte, resulting in the well-known GC equation [19]:

$$1 = \int_0^{t^{(p)}_{r_{mod},i}} \frac{dt}{t_M(t)\cdot[1 + k_i(t)]} (p = 1 \ldots N_{Exp}) \tag{6}$$

where: N_{Exp}: number of temperature programs tested; $t^{(p)}_{r_{mod},i}$: Retention time predicted by the model in the case of the pth temperature program for the ith compound, where $p = 1 \ldots N_{Exp}$ and $i = 1 \ldots N_{comp}$ (N_{comp} is the number of compounds).

3. Experimental procedure

Retention time data from a series of alkanes for different temperature programs were collected for this study. Analyzed alkanes were: Octane Nonane, Decane, Dodecane, Tetradecane and Pentadecane. All temperature programs started with an isotherm at 50 °C lasting 3 min, followed by a ramp with a heating rate ranging between 5 and 40 °C/min until reaching 175 °C. The last step, common to all temperature programs, consisted of an isotherm at 175 °C lasting 5 min. The data collected is shown in Table 1.

The experiments were performed on a Shimadzu GC-2010 Plus gas chromatograph, equipped with a FID-2010 Plus detector, DB-5 column (Agilent) with length $L = 30$ m, internal diameter $d_i = 250$ μm and film thickness $d_f = 0.25$ μm. Inlet and outlet pressures were kept fixed at 2.013×10^5 and 1.013×10^5 [Pa], respectively, where the carrier gas used was N_2.

In order to determine t_M, experiments were performed with dichloromethane [20] at various temperatures in the range between 150 °C and 275 °C, as described in a previous work [21]. To the measured t_M values, an equation correlating t_M to the viscosity of the carrier gas used at each used temperature was adjusted – as shown in Eq. (7).

$$t_M\,[\mathrm{s}] = 4.80037 \times 10^6 \cdot \mu\,[\mathrm{Pa}\cdot\mathrm{s}] + 13.9052 \tag{7}$$

4. Parameter estimation

Estimation of parameters present in Eq. (5) consists in determining the values of ΔH and ΔS (and ΔC_p) that minimize the error between retention times predicted by the model (Eq. (6)) and the experimental ones. Such problem corresponds to a non-linear regression in the non-isothermal case and, thus, must be solved iteratively, using a method to optimize the performance index set.

Although different performance indices may be used to estimate parameters such as, for example, the maximum error in absolute value and the sum of errors in absolute value, among others, the most common is the application of a quadratic criterion, such as the sum of the square error (*SSE*). The minimization of SSE implies the maximization of the coefficient of determination R^2.

$$SSE = \sum_{p=1}^{N_{exp}} \left[t^{(p)}_{r_{exp},i} - t^{(p)}_{r_{mod},i}(\Delta H, \Delta S, \Delta C_p, T) \right]^2 \tag{8}$$

Table 1
Experimental retention times for an alkane series with different program temperatures.

t_r [min]	Program temperature [N_{Exp}]							
	1	2	3	4	5	6	7	8
	Heating rate [°C/min]							
	5	10	15	20	25	30	35	40
Octane	5.014	4.821	4.688	4.586	4.505	4.442	4.386	4.337
Nonane	7.854	6.925	6.408	6.081	5.846	5.680	5.545	5.435
Decane	11.178	8.990	7.971	7.367	6.967	6.682	6.467	6.311
Dodecane	17.643	12.560	10.502	9.365	8.685	8.232	7.907	7.663
Tetradecane	23.475	15.619	12.786	11.370	10.518	9.952	9.553	9.248
Pentadecane	26.153	17.231	14.211	12.705	11.799	11.195	10.772	10.444

where, $t^{(p)}_{rexp,i}$: measured retention time of the pth temperature program, obtained experimentally for the ith compound.

In an analysis, it is expected, or at least desirable, that experimental retention times of components be of the order of minutes, though depending on the combination of parameters used the model might be able to predict retention times of the order of days. Clearly, in this situation, *SSE* would assume a very high value. As will be shown in section 4.1.1, if the *SSE* ($ln(SSE)$) logarithm be considered, this will result in a tilt-up flat surface in the high retentions region.

At the other end, certain combinations of parameters can result in a very low retention, even for the lowest temperature used in the experiments. In the latter case, the retention time predicted by the model for any temperature program tends to the value of the mobile phase (t_M), that is, without retention.

Thus, in parameter estimation, it is desirable to discriminate the region within the search space with the greatest possibility for finding the global optimum from those that represent bad combinations. Thus the domain of the objective function has been identified, and divided into different regions.

4.1. Objective function domain division in different regions

In order to graphically display the different regions of the $ln(SSE)$, it was initially considered that it depended only on ΔH and ΔS; nevertheless, the same regions will exist if additional parameter ΔC_p is used, as shown in Section 5.

From the present study it was determined that due to ΔH and ΔS, the *SSE* surface can be divided into different regions; it was the compound *dodecane* that which was used to illustrate that the objective function domain is divided into three well-defined regions – namely:

Region S1: the parameter combinations contained in this region will predict very high retention even at high temperatures;

Region S2: where the global optimum or the solution to the problem is contained;

Region S3: combinations of parameters contained in this region will predict very small retention even at low temperatures.

It can be shown that regions S1 and S3 will have a predictable and well behaved conduct, regardless of the component to which they relate.

Fig. 1 shows the surface of the $ln(SSE)$ in the case of dodecane using the data shown in Table 1. Parameters ΔH and ΔS were varied in the range of −200 and 0 [kJ/mol] and −200 and 0 [J/(mol·K)], respectively. It is worth noting that the application of the natural logarithm to the *SSE* was done solely to assist with the visualization, more specifically to mitigate the high values recorded in the S1 region; notwithstanding, such a transformation does not change the location of the global optimum.

In Fig. 1 the defined regions S1, S2 and S3 can be observed. Region S1 displays an inclined plane behavior, and S3 tends to be a constant. Region S2, which is intermediate and located between S1 and S3, generally displays a small dimension. An analytical expression of the behavior of $ln(SSE)$ in the case of regions S1 and S3 can be derived, and it is described in Sections 4.1.1 and 4.1.2, respectively. Obviously, there are transition zones between regions S1 and S2, and between regions S2 and S3; however, this does not change the format of $ln(SSE)$ that is in Fig. 1. The same behavior is observed for other tested compounds of the series of alkanes.

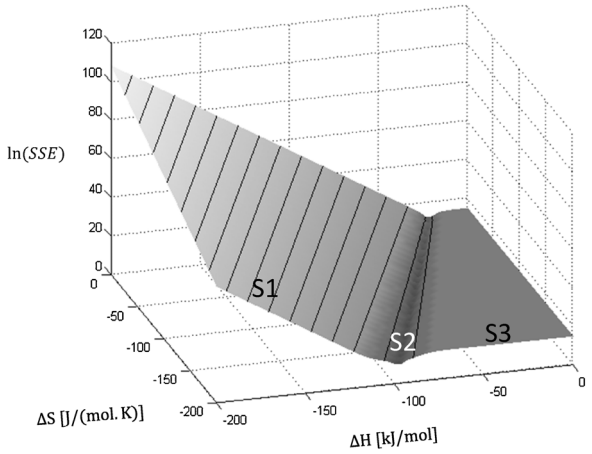

Fig. 1. S1, S2 and S3 regions on ln(SSE) surface for dodecane.

The reason for dividing the estimation problem into regions is that the retention factor $k(T)$ present in the retention time model (Eq. (5)) can vary by orders of magnitude depending on the temperature value. Because of the exponential relationship between $k(T)$ and temperature, the retention factor can assume values far below or above the unit.

It is important to note that the regions described are not specific to dodecane and were observed for all the alkanes that were described in Table 1. In the deductions made to elucidate the behaviors of regions S1 and S3, no simplifying hypothesis related to the chemical nature of the compounds was used. Thus, the division of the objective function in regions S1, S2 and S3 will occur regardless of the analyte and of the stationary phase considered – and this is characteristic of the problem under study.

The presence of regions S1 and S3 makes optimization more difficult, and turns parameter estimation into a non-linear problem, for S1 and S3 are regions with high and low sensitivity, respectively, with respect to parameters ΔH and ΔS.

In the case of region S1, the exponential behavior, which is monotonic, may indicate the optimization method, which is an easily solvable problem. In general, in such a situation a calculus-based method tends to increase the displacement step. As a result, a point initialized in region S1 can evolve quickly and jump directly to region S3, ignoring the existence of region S2. In the following Sections 4.1.1 and 4.1.2 mathematical expressions that represent

the regions S1 and S3 in terms of the parameters retention time model will be derived.

Considering region S3, if a method based on derivatives is used for estimating parameters, it will be difficult for convergence to happen, because the *SSE* displays constant behavior, therefore there is no information as to which direction is going to reduce the *SSE*.

4.1.1. Region S1

In the case of region S1 retention times predicted by the model will be significantly larger than the experimental values. Thus, the value of *SSE* in Eq. (8) may be approximated as shown in Eq. (9):

$$SSE = \sum_{p=1}^{N_{exp}} \left[t^{(p)}_{r_{exp},i} - t^{(p)}_{r_{mod},i}(\Delta H, \Delta S, T) \right]^2 \approx \sum_{p=1}^{N_{exp}} \left[t^{(p)}_{r_{mod},i}(\Delta H, \Delta S, T) \right]^2 \tag{9}$$

So as to obtain an expression for the SSE and ln(SSE), at first the following must be defined:T_{max}: it is assumed that all temperature programs used will finalize at a single temperature. Given that programs show increasing monotonic behavior over time, said temperature will be the highest one applied;$t_{T_{max}}$: the longest time it takes to reach maximum temperature (T_{max}:) amongst all temperature p-programs (at estimation). Mathematically, $t_{T_{max}}$ was defined according to Eq. (10). As an example, considering that the temperature programs used offer different heating rates, $t_{T_{max}}$ would be the time the program with the lowest heating rate takes to reach T_{max}.

$$t_{T_{max}} = \max_{p=1\ldots N_{exp}} (t^{(p)}_{T_{max}}) \tag{10}$$

Subsequently, the model given by Eq. (6) must be evaluated, and split into two integration intervals. The first interval considers the integration range between zero and $t_{T_{max}}$, and the second one between $t_{T_{max}}$ and retention time (see Eq. (11)):

$$1 = \left(\frac{x = L}{L}\right) = \int_0^{t_{T_{max}}} \frac{dt}{t_M(t)\cdot[1 + k_i(t)]} + \int_{t_{T_{max}}}^{t^{(p)}_{r,i}} \frac{dt}{t_M(t)\cdot[1 + k_i(t)]} \quad (p=1\ldots N_{Exp}) \tag{11}$$

The value of $t_{T_{max}}$ is, usually, of the order of minutes. At region S1 it is assumed that the values of $t^{(p)}_{r_{mod},i}$ are higher than $t_{T_{max}}$.

In Eq. (11), the unit value was written conveniently as $(x = L)/L$ to remind that the terms of the right side represent normalized distances, that is, the length run by the analyte divided by the column length. Usually it is desirable to determine the time that an analyte takes to exit the column, thus the unit value is shown directly.

Treating the integrals present on the right side of Eq. (11) as normalized distances, the parameters of the model in region S1 are such that the retention time can be very large (of the order of days). Thus, the length run by an analyte between zero and $t_{T_{max}}$ represents a small value when compared to the length of the column L and, therefore, the first integral contribution may be neglected in Eq. (11), which results in Eq. (12):

$$1 = \int_{t_{T_{max}}}^{t^{(p)}_{r,i}} \frac{dt}{t_M(t)\cdot[1 + k_i(t)]} (i = 1\ldots NC) \tag{12}$$

Assuming that T is constant and equal to T_{max} for times longer than $t_{T_{max}}$, Eq. (12) presents an analytical solution for $t^{(p)}_{r,i}$ according to Eq. (13).

$$t^{(p)}_{r,i} = t_{T_{max}} + t_M(T_{max})\cdot[1 + k_i(T_{max})] \quad (p = 1\ldots N_{exp}) \tag{13}$$

Again, seeing as retention time is too high, the unit in relation to the value of $k(T_{max})$ can be neglected, as can the $t_{T_{max}}$ contribution from Eq. (13), which results in Eq. (14):

$$t^{(p)}_{r,i} = t_M(T_{max})\cdot k(T_{max}) \quad (p = 1\ldots N_{exp}) \tag{14}$$

By replacing Eq. (14) in Eq. (9) and applying the natural logarithm to the latter, an expression for *ln(SSE)* for the region S1 is obtained:

$$\begin{aligned} ln(SSE) &= ln\left\{\sum_{i=1}^{N_{exp}} [t_M(T_{max})\cdot k(T_{max})]^2\right\} = ln\{N_{exp}\cdot[t_M(T_{max})\cdot k(T_{max})]^2\} \\ ln(SSE) &= ln\{N_{exp}\} + 2ln\{t_M(T_{max})\} + 2ln(k(T_{max})) \\ ln(SSE) &= ln\{N_{exp}\} + 2ln\{t_M(T_{max})\} + 2\left(-ln\beta - \frac{\Delta H}{RT_{max}} + \frac{\Delta S}{R}\right) \\ ln(SSE) &= \underbrace{ln\{N_{exp}\} + 2ln\{t_M(T_{max})\} - 2ln\beta}_{a1} - \underbrace{\frac{2}{RT_{max}}}_{a2}\Delta H + \underbrace{\frac{2}{R}}_{a3}\Delta S \\ ln(SSE) &= a_1 - a_2\cdot\Delta H + a_3\cdot\Delta S \end{aligned} \tag{15}$$

The result shown in Eq. (15) is that the *ln(SSE)* behaves in the S1 region according to an inclined plane due to variables ΔH and ΔS, or that *SSE* depends exponentially on parameters ΔH and ΔS.

It is worth stressing that the assumption that all temperature programs will finish at $T = T_{max}$ was only made for the purposes of obtaining a simplified analytical result: SSE depends exponentially on parameters ΔH and ΔS from the S1 region. However, if this assumption is not made, this would not change the fact that the error criterion presents high sensitivity in relation to variations of parameters ΔH and ΔS on region S1.

The application of the natural logarithm in Eq. (8) does not modify the original problem of parameter estimation, seeing as this function presents a monotonic behavior with respect to its argument. This transformation was applies solely to facilitate understanding of the behavior of region S1.

4.1.2. Region S3

In this region the combinations of parameters ΔH and ΔS will result in retention times very close to the t_M of the mobile phase. Because t_M does not depend on these parameters, this implies that the surface of *SSE* and that of its logarithm will tend to a constant value in region S3, as shown in Eq. (16):

$$SSE = \sum_{i=1}^{N_{exp}} [t^{(p)}_{r_{exp},i} - t^{(p)}_{r_{mod},i}(\Delta H, \Delta S, T)]^2 \approx \sum_{i=1}^{N_{exp}} [t^{(p)}_{r_{exp},i} - t_M(t)]^2 \tag{16}$$

It is worth noting that any temperature programming will present the same retention time for parameters located in region S3, as can be seen in Eq. (16). Thus, flat behavior in the *log* scale will occur not only for *SSE*, but also for other functions goal such as, for example, the sum of absolute error, or errors greater in absolute value among all the temperature programs.

Region S2 represents a transition zone between S1 and S3, and it is the region where the global optimum is.

4.2. Algorithm for discriminating different regions of parameter estimation (ΔH, ΔS)

From what was set forth in 4.1, it would be desirable, for parameters estimation, to determine the region within the search space where there is a higher possibility for the global optimum location,

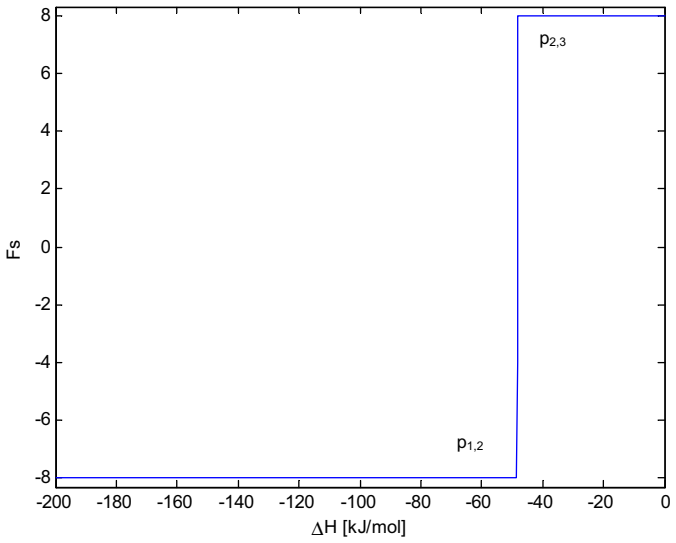

Fig. 2. *Fs* function for a fixed ΔS −70 J/(mol·K).

in this case, region S2. For this study, said goal was achieved from the definition of the *Fs* function, according to Eq. (17):

$$Fs(\Delta H, \Delta S) = \sum_{p=1}^{N_{exp}} sgn(t^{(p)}_{r_{exp},i} - t^{(p)}_{r_{mod},i}) \quad (17)$$

Where sign function, written as sgn(x) is defined as Eq. (18):

$$sgn(x) = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases} \quad (18)$$

Numerically, the *Fs* function is equal to Eq. (19):

$$Fs(\Delta H, \Delta S) = Np(t^{(p)}_{r_{exp},i} > t^{(p)}_{r_{mod},i}) - Np(t^{(p)}_{r_{exp},i} < t^{(p)}_{r_{mod},i}) \quad (19)$$

where N_P indicates the number of points that satisfy each inequality.

The *Fs* function will assume a maximum value equal to N_{exp} when all retention times predicted by the model are below the experimental data. The minimum possible value assumed by *Fs* (same as $-N_{exp}$) will occur when the retention times predicted by the model are above those of the experimental data. The first case occurs if the retention factor is below what it should be; and the last if the retention factor is too high.

The *Fs* function is discontinuous and takes integers values only, as it represents the difference between the number of points, predicted by the model, which are below and above the experimental data.

Shown in Fig. 2 are the values assumed by the *Fs* function, considering $\Delta S = -70$ J/(mol·K) and ΔH varying between −200 and 0 kJ/mol, in the case of the dodecane, and utilizing temperature programs described in Table 1. The values of $p_{1,2}$ and $p_{2,3}$ correspond to transition points between regions S1 and S2, and S2 and S3 respectively, so as the example assumes the following values:

$p_{1,2}$: ($\Delta H = -48.4484$ kJ/mol; $\Delta S = -70$ J(mol·K); $Fs = -8$)

$p_{2,3}$: ($\Delta H = -48.0480$ kJ/mol; $\Delta S = -70$ J(mol·K); $Fs = 8$)

In Fig. 2 the *Fs* function presents a monotonically crescent behavior, as the value of ΔH increases and assumes values between $-N_{Exp}$ and N_{Exp}. The transition zone with values higher than $-N_{Exp}$ and lower than N_{Exp} serves as a good approximation for the location of S2.

The same behavior displayed in Fig. 2 occurs for other ΔS values; the only change is the location of the transition region between $-N_{Exp}$ and N_{Exp}. Thus, regions S1, S2 and S3 can be discriminated in terms of the values as given by:

Region S1: $Fs = -N_{Exp}$

Region S2: $-N_{Exp} < Fs < N_{Exp}$

Region S3: $Fs = N_{Exp}$

For a given ΔS, the limit between regions S1 and S2 can be determined as the highest value of ΔH for which ($Fs = -N_{Exp}$), that is, point $p_{1,2}$ can be calculated as being the highest inferior limit.

The limit between regions S2 and S3 can be determined as the lowest value of ΔH, for which ($Fs = N_{Exp}$), thus point $p_{2,3}$ can be calculated as the lowest superior limit.

It is worth noting that $p_{1,2}(\Delta S)$ and $p_{2,3}(\Delta S)$ are located in *Fs* discontinuities. A useful device for determining $p_{1,2}$ is to add a constant to *Fs* in a way that the modified *Fs* function shows contrary signs to the left and right of $p_{1,2}$. Thus, $p_{1,2}$ can be determined, for each ΔS, by a root finding method, such as Bisection [22]. A similar procedure can be applied when determining $p_{2,3}$, except the modified *Fs* function must be obtained be decreasing *Fs* from a constant.

The modified *Fs* functions $F_{1,2}$ and $F_{2,3}$ are defined in Eqs. (20) and (21), for determining $p_{1,2}$ and $p_{2,3}$ respectively. The (N_{Exp}–0.5) constant is not arbitrary and provokes sign change between regions S1 and S2 (in the case of $F_{1,2}$) and regions S2 and S3 (in the case of $F_{2,3}$).

$$F_{1,2} = Fs + (N_{Exp} - 0.5) \quad (20)$$

$$F_{2,3} = Fs - (N_{Exp} - 0.5) \quad (21)$$

As an example, in the case of $F_{2,3}$, the maximum possible value will be equal to 0.5 (to the right of $p_{2,3}$), for the highest assumed value by *Fs* is equal to N_{Exp}. As *Fs* varies discontinuously (in whole values), the same will happen with $F_{2,3}$, wherein to the left of $p_{2,3}$ negative values for $F_{2,3}$ will be observed.

By allowing ΔS to vary in the range of interest (in this case between −200 and 0 J/mol K), functions $F_{1,2}(\Delta S)$ and $F_{2,3}(\Delta S)$ will be obtained, from which the limits between regions S1, S2 and S3 will be determined.

The Bisection method must be initialized as an interval in which the research function displays contrary signs at the extremities. At each iteration of the method the interval is reduced by a factor of two, and after 10 evaluations the search interval is decreased by 2^{10} (approximately 1000 times) from the original. Thus, a good approximation for $p_{1,2}(\Delta S)$ can be obtained with a small number of evaluations of the integrand. The same procedure must be repeated for determining $p_{2,3}(\Delta S)$.

In Fig. 3 the results of applying the algorithm described for determining region S2 for the case of dodecane are shown. As can be observed, region S2 is small compared to S1 and S3.

Once more, it is important to stress that the regions described in Fig. 3 are not specific to the dodecane, when they were observed for all alkanes. Additionally, such regions should occur for any component, irrespective of its chemical nature. An alternative proof of the existence of such regions (S1, S2, and S3), for any analyte regardless of the stationary phase, can be found in Appendix A.

It should be highlighted that, usually, the Bisection method is applied for determining a root contained in an interval; however, for this study, that method was used for determining points $p_{1,2}$ and $p_{2,3}$ which compose an interval of small dimensions whose extremities display contrary signs in relation to the type of *Fs* function considered. This way, by applying Bisection, points for which the research function assumes value zero are not sought, which is the usual application for the method. In the case of $p_{1,2}$ and $p_{2,3}$, value zero may not even be contained in the set image of these functions.

Beyond determining the boundaries between regions S1, S2 and S3, the *Fs* function for determining a starting point (initial guess) for parameter estimations can be used. In this case, it suffices to

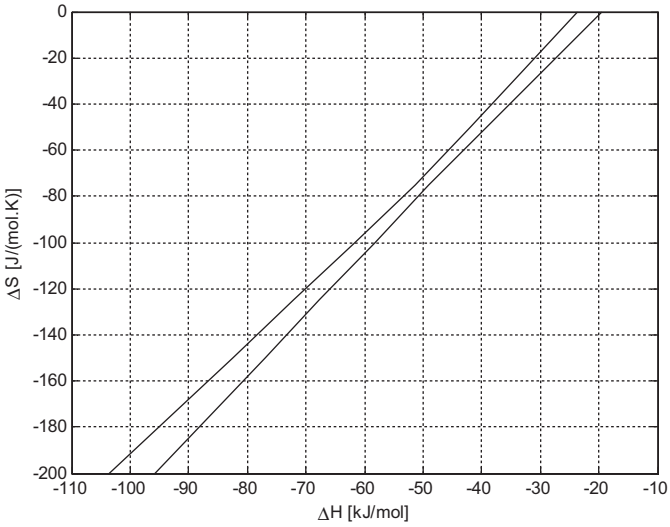

Fig. 3. Estimate of S2 region for dodecane.

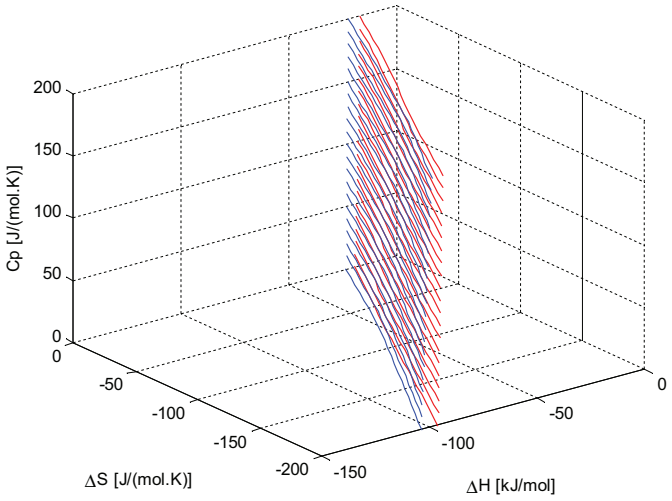

Fig. 4. Estimate of S2 region for dodecane (three parameters).

arbitrate a ΔS value in the range of interest and apply the Bisection method to the original *Fs* to find the value of the respective ΔH contained in S2.

The determination of a good initial guess can be improved by evaluating, simultaneously to the application of the bisection method, the value of the *SSE* criterion and keeping the minimum value solution found. This does not mean a significant increase in computational time, the most demanding step is the resolution of the model for predicting t_r and not the calculation of *Fs* and *SSE*.

Additionally, the initialization described above can be repeated a few times for different initial condition and keep the best solution obtained (in terms of the smallest value of the error criterion) in all replications. The latter procedure will be called, throughout this work (a two-part series), *specialized initialization*.

As defined, the specialized initialization has two parameters: the number of times the search method will be applied for different initial conditions, and the corresponding stopping criterion, that is, the maximum number of evaluations of the model permitted.

5. Application of the region discrimination algorithm to three parameters (ΔH, ΔS, ΔC_p)

The methodology for parameter estimation was displayed in Section 4 considering only two parameters (ΔH and ΔS) to facilitate visualization. However, it is common to consider ΔH and ΔS dependency to temperature according to Eqs. (1)–(5). In this case there is an additional parameter to be estimated: ΔC_p.

In the case of the problem with three parameters, the existence of regions remains unchanged. Regions S1, S2 and S3 described above continue to exist. In fact, parameter ΔC_p is of secondary importance when compared to ΔH and ΔS, and it is disregarded in many works available from the literature [8,19,23].

The main change, should parameter ΔC_p be introduced, is that the *Fs* function set as $Fs(\Delta H, \Delta S)$ will be set as $Fs(\Delta H, \Delta S, \Delta C_p)$. As a result, the modified *Fs* functions should also be updated to include the value of ΔC_p.

Fig. 4 displays the results of the application of the methodology for determining region S2 in the case of three parameters. In the same manner as it occurred for two parameters, it can be observed that S2 is small in relation to S1 and S3.

Fig. 4 was designed by initially defining a mesh of points in terms of ΔS and ΔC_p, wherein the Bisection method was applied for finding the values of ΔH that made $F_{1,2}$ and $F_{2,3}$ change sign.

In the next article of this series the effect of the specialized (robust) initialization for parameter estimation in GC will be shown; a comparison with the purely random initialization will be run as well.

6. Conclusions

As shown in this study, in the case of the problem for parameter estimation for prediction of retention time in GC, the surface of the approximation error criterion obtained due to thermodynamic parameters (ΔH, ΔS, ΔC_p) can be split into well-defined regions: (a) Where combination of parameters will result in high retention even for high temperatures; (b) The global optimum or solution to the problem is contained in this region; (c) Combinations of parameters contained in this region will result in very low retention even for low temperatures.

In the deductions made to elucidate the behavior of the described regions, no simplifying hypothesis related to the chemical nature of the compounds was used. Thus, the division of the surface of the objective function in regions will occur independently of the component considered, which is characteristic of the problem under study.

The presence of three main regions occurs whether parameter estimations are done with two or three parameters (ΔH and ΔS) and (ΔH, ΔS, ΔC_p), respectively.

The same behavior in terms of regions can be observed for different performance criteria (norms) based on the sum of the square error, the maximum percentage error and the maximum absolute error.

An algorithm was developed to discriminate the different regions of the error criterion. Its main utility is the robust initialization for parameter estimation.

Conflict of interest

The authors confirm that there are no conflicts of interest regarding this paper.

Acknowledgments

The authors thank CAPES, CNPq for their financial support. A. Wüst Zibetti thanks to CAPES-PNPD (2014/2015) for their assistantship.

Appendix A. Alternative proof of the existence of three regions in parameter estimation

An alternative proof of the existence of three regions – already presented, and denominated S1, S2 and S3 – will be further demonstrated in this appendix. The main objective of this extension is to prove that the error surface regions exist, regardless of the analyte and stationary phase. In the deductions proposed – be it the one found in Section 4 or this one – no simplifying hypothesis of the model was assumed; that is, it was not assumed that the analyte is of any specific kind (e.g.: structural form or carbon chain length). Another important point is that this proof is valid for any stationary phase. This is because the interaction between the analyte and the stationary phase appears precisely in parameters (ΔH, ΔS and ΔC_p), estimated from the retention factor equation.

To demonstrate evidence in a simplified form, only parameters (ΔH and ΔS) were taken into account. Thus, in order to facilitate visualization of the separation between regions S1, S2 and S3, ΔC_p was not considered for this alternative proof.

Mathematical and physical proof come from the retention factor model $k(T)$, where a high value of k corresponds to high retention, and a low value of k corresponds to low retention.

Given that the expression for the retention factor due to temperature, thermodynamic parameters and volume ratio of the phases is given by Eq. (A-1),

$$\ln(k) = -\frac{\Delta H}{RT} + \frac{\Delta S}{R} - \ln(\beta) \tag{A-1}$$

consider the case of a hypothetical component (analyte) that presents small retention (e.g.: $k \leq 0.01$) in a given column, even for the lowest temperature ($T = T_{min}$) on the program: by isolating ΔS from Eq. (A-1), the result is an inequality, named D_A, as shown in Eqs. (A-2)–(A-4).

$$\ln(k) = -\frac{\Delta H}{RT_{\min}} + \frac{\Delta S}{R} - \ln(\beta) \leq \ln(0.01) \tag{A-2}$$

$$\Delta S \leq \frac{\Delta H}{T_{\min}} + \underbrace{R[\ln(\beta) + \ln(0.01)]}_{c_{\min}} \tag{A-3}$$

$$\Delta S \leq \frac{\Delta H}{T_{\min}} + c_{\min} \quad (\text{inequality } D_A) \tag{A-4}$$

The meaning of inequality D_A is that any combination of parameters that satisfies it will imply that the hypothetical component associated presents negligible retention, and therefore moves similarly to how it does in the mobile phase.

In another extreme case, the case of a hypothetical component (analyte) which presents high retention (e.g.: $k \geq 100$) is considered, even for the highest temperature ($T = T_{max}$) on the program. Again, isolating ΔS results in inequality D_B, as shown in Eqs. (A-5)–(A-7).

$$\ln(k) = -\frac{\Delta H}{RT_{\max}} + \frac{\Delta S}{R} - \ln(\beta) \geq \ln(100) \tag{A-5}$$

$$\Delta S \geq \frac{\Delta H}{T_{\max}} + \underbrace{R\left[\ln(\beta) + \ln(100)\right]}_{c_{\max}} \tag{A-6}$$

$$\Delta S \geq \frac{\Delta H}{T_{\max}} + c_{\max} \, (\text{inequality } D_B) \tag{A-7}$$

The interpretation for inequality D_B is: any combination of parameters that satisfies it will imply that the hypothetical associated component will display very high retention, even at high temperatures. Such compound would probably remain retained in the column, or its analysis would require an excessively large amount of time, hence this column would not be suitable for the analysis of such component.

Inequalities D_A and D_B do not intersect, at least not in the region of interest, that is, for negative values of parameters ΔH and ΔS, which are associated with the transition of an analyte from the mobile phase (vapor) to the stationary phase (liquid).

Whether equations Eqs. (A-8) and (A-9) – called equality E_A and E_B – are applicable, respectively, at the limit of the validity of inequalities D_A and D_B, that means $D_A \leq E_A$ and $E_B \leq D_B$ for any predicted value of ΔH considered.

$$\Delta S = \frac{\Delta H}{T_{min}} + c_{min} \, (\text{Equality } E_A) \tag{A-8}$$

$$\Delta S = \frac{\Delta H}{T_{max}} + c_{max} \, (\text{Equality } E_B) \tag{A-9}$$

In Eqs. (A-8) and (A-9), the linear coefficient (intercept) from Eq. (A-8) is smaller than its corresponding parameter for Eq. (A-9) ($c_{min} < c_{max}$). Additionally, the angular coefficient (slope) from Eq. (A-8) is higher than that of Eq. (A-9) – that is, ($1/T_{min} > 1/T_{max}$), which implies:

$$c_{min} < c_{max} \tag{A-10}$$

$$\frac{1}{T_{min}} > \frac{1}{T_{max}} \tag{A-11}$$

$$\frac{\Delta H}{T_{min}} < \frac{\Delta H}{T_{max}} \, (\text{valid for negative } \Delta H) \tag{A-12}$$

Adding inequalities Eqs. (A-10) and (A-12), the following happens:

$$\underbrace{\frac{\Delta H}{T_{\min}} + c_{\min}}_{E_A} < \underbrace{\frac{\Delta H}{T_{\max}} + c_{\max}}_{E_B} \tag{A-13}$$

where E_A and E_B are the lines that define the two regions (of low or no retention, and high retention, respectively). The two extreme regions, which are divided by the lines E_A and E_B, are called: D_A – the region where any combination of parameters ΔH and ΔS results in little or no retention; and D_B – high retention region, for any combination of parameters. These parameters belong to the region of physical interest, that is, negative values of ΔH and ΔS.

Combining the result obtained from Eq. (A-13) with the fact that $D_A \leq E_A$ and $E_B \leq D_B$, it follows that:

$$D_A \leq E_A < E_B \leq D_B \tag{A-14}$$

$$D_A < D_B \tag{A-15}$$

Thus, one can divide the relationship between ΔS and ΔH into three regions: D_A, D_B, and the area between them. The region between the two inequalities (where D_A and D_B are false) comprises the possible combinations of parameter values that show minimal retention. That is, this region (between D_A and D_B) is the one that represents the feasible combination of parameters which depicts the analyte moving within the column at a speed lower than that of the mobile phase – without resulting in extreme retention, however, so as the analyte does not become adsorbed on the stationary phase.

As an example, Fig. A1 shows an illustration with the following characteristics:

- In the case of negligible retention limit ($k = 0.01$);
- In the case of high retention limit ($k = 100$);
- Relationship between the volume of the phases in the column: $\beta = 250$ [values may vary between ≈60 and ≈500]
- $T_{min} = 50\,^{\circ}C$
- $T_{max} = 250\,^{\circ}C$

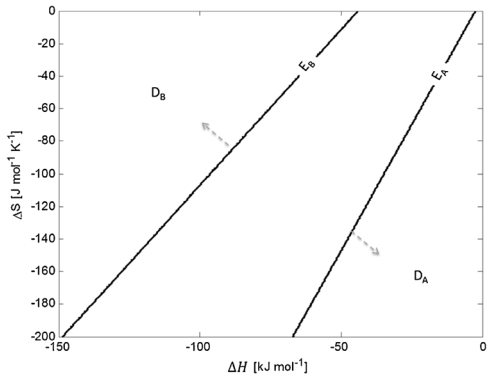

Fig. A1. Illustration of the different regions of interest for optimization.

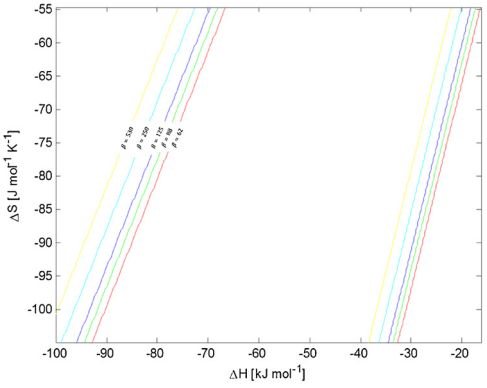

Fig. A2. Illustration of the different regions of interest for optimization, according to β values: 62, 88, 125, 250 and 530, for T_{min} = 50 °C and k = 0.01, T_{max} = 250 °C and k = 100.

An analogy between the regions shown in Fig. A1 and Fig. 3 can be made:

Fig. 3	Fig. A1
S1	Region in which D_B can be verified (upper left)
S2	Region in which D_A and D_B are false
S3	Region in which D_A can be verified (bottom right)

Comparing the size of the region between D_A and D_B (Fig. A1) with that of region S2 (Fig. 3), it is noteworthy that the former is too large, as this is a more conservative deduction – that is, it considers the extreme limits of hypothetical components, be them retained or not. Additionally, in Fig. 3, the S2 region has been particularized for a specific component: Dodecane (β = 250), that is, this is a very specific case. In Fig. A2 equalities E_A (T_{min} = 50 °C and k = 0.01) and E_B (T_{max} = 250 °C and k = 100), according to different β values between 62 and 530, are shown. β values are equivalent to geometric data from commonly marketed columns. It is important to note that changes in β values cause a shift in both the intercepts. The higher the β values, the higher the intercepts of equalities (E_A and E_B) (Fig. A2).

As noted, no restrictions regarding the combinations of possible values for ΔH and ΔS were taken into account for the deduction. Again, seeing as the information about the interaction between analyte and stationary phase influences the model for predicting retention time through the parameters, it is believed that the deduction applies to any type of analyte and stationary phase.

Additionally, no restriction to the temperature program applied took place, besides de obvious one (temperature of a minimum value T_{min}, is raised until it reaches maximum value T_{max}).

It is worth stressing that the proof displayed in Appendix A is simpler than the one shown in Section 4.1. Nonetheless, its results reinforce the general nature of the claim that these regions exist for

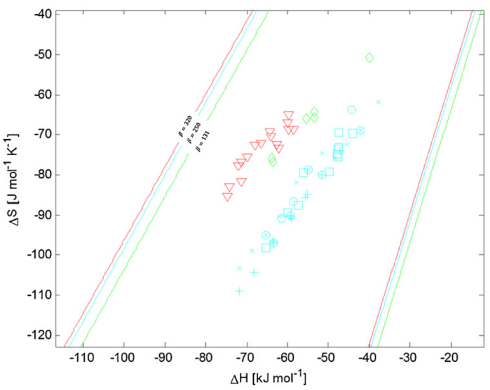

Fig. A3. Region S2 generalized according to the evidence presented in Appendix A. Analytes: □ – Grob mix (column SLB5ms, β = 205) [24]; X – n-alkane mix (column Supelco Wax, β = 250)[11]; + – n-alkane mix (column SLB5ms, β = 250) [11]; ◯ – Dodecanone, 1 Dodecanol and 2-Dodecanone (in columns SLB5ms, SPB50 and Supelco Wax β = 250) [11]; ◇ – different analytes obtained by Dose – Table I – [20] (column D_B-17, β = 131); ∇ – different analytes obtained by Dose–Table II – [20] (column D_B-210, β = 320). Lines have been represented with the following characteristics: T_{min} = 30 °C and k = 0.01, T_{max} = 275 °C and k = 100 and respective β values β = 131, 250 and 320.

any compound, regardless of their stationary phase(s). This proof demonstrates, more conservatively, the existence of what has been called region S2 inside the region that is separated by the extreme cases of low and high retention. So as to illustrate this contribution, Fig. A3 displays several analytes parameterized in different stationary phases (information obtained from some works in the literature, namely [11,20,24]).

As it can be seen in Fig. A3, all analytes have their respective parameters (ΔH and ΔS) within the generalized region from this proof. It should be noted that parameter ΔC_p adds a coordinate to the graph; thus, what is demonstrated via Fig. A3 doubles as experimental validation of the existence of a region that is independent from the type of analyte and of the stationary phase.

References

[1] F. Aldaeus, Y. Thewalim, A. Colmsjö, Prediction of retention times and peak widths in temperature-programmed gas chromatography using the finite element method, J. Chromatogr. A 1216 (2009) 134–139, http://dx.doi.org/10.1016/j.chroma.2008.11.038.
[2] S. Vezzani, P. Moretti, G. Castello, Classification and comparison of capillary columns by determination of the solution enthalpy of polar and non polar probes, J. Chromatogr. A 1101 (2006) 261–267, http://dx.doi.org/10.1016/j.chroma.2005.10.006.
[3] H. Snijders, H.-G. Janssen, C. Cramers, Optimization of temperature-programmed gas chromatographic separations II. Off-line simplex optimization and column selection, J. Chromatogr. A 756 (1996) 175–183, http://dx.doi.org/10.1016/S0021-9673(96)00626-7.
[4] T.M. McGinitie, B.R. Karolat, C. Whale, J.J. Harynuk, Influence of carrier gas on the prediction of gas chromatographic retention times based on thermodynamic parameters, J. Chromatogr. A 1218 (2011) 3241–3246, http://dx.doi.org/10.1016/j.chroma.2010.09.068.
[5] B. Karolat, J. Harynuk, Prediction of gas chromatographic retention time via an additive thermodynamic model, J. Chromatogr. A 1217 (2010) 4862–4867, http://dx.doi.org/10.1016/j.chroma.2010.05.037.
[6] T.M. McGinitie, J.J. Harynuk, Prediction of retention times in comprehensive two-dimensional gas chromatography using thermodynamic models, J. Chromatogr. A 1255 (2012) 184–189, http://dx.doi.org/10.1016/j.chroma.2012.02.023.
[7] X. Li, G. Fan, C. Gong, M. Ao, H. Li, Prediction of retention times in temperature programmed gas chromatography using the retention equation derived from crystallization behavior of polymer, J. Chromatogr. A 1277 (2013) 76–83, http://dx.doi.org/10.1016/j.chroma.2012.12.051.
[8] F. Aldaeus, Y. Thewalim, A. Colmsjö, Prediction of retention times of polycyclic aromatic hydrocarbons and n-alkanes in temperature-programmed gas chromatography, Anal. Bioanal. Chem. 389 (2007) 941–950, http://dx.doi.org/10.1007/s00216-007-1528-0.
[9] B. Peng, M.-Y. Kuo, P. Yang, J.T. Hewitt, P.G. Boswell, A practical methodology to measure unbiased gas chromatographic retention factor vs. temperature

relationships, J. Chromatogr. A 1374 (2014) 207–215, http://dx.doi.org/10.1016/j.chroma.2014.11.018.
[10] P.G. Boswell, P.W. Carr, J.D. Cohen, A.D. Hegeman, Easy and accurate calculation of programmed temperature gas chromatographic retention times by back-calculation of temperature and hold-up time profiles, J. Chromatogr. A 1263 (2012) 179–188, http://dx.doi.org/10.1016/j.chroma.2012.09.048.
[11] T.M. McGinitie, H. Ebrahimi-Najafabadi, J.J. Harynuk, Rapid determination of thermodynamic parameters from one-dimensional programmed-temperature gas chromatography for use in retention time prediction in comprehensive multidimensional chromatography, J. Chromatogr. A 1325 (2014) 204–212, http://dx.doi.org/10.1016/j.chroma.2013.12.008.
[12] D.E. Goldberg, Genetic Algorithms, Pearson Education, 2006.
[13] M.S. Bazaraa, H.D. Sherali, C.M. Shetty, Nonlinear Programming: Theory and Algorithms, John Wiley & Sons, 2013.
[14] T.F. Edgar, D.M. Himmelblau, L.S. Lasdon, Optimization of chemical processes, 2001.
[15] P.J.M. van Laarhoven, E.H.L. Aarts, Simulated Annealing: Theory and Applications, Springer, Netherlands, Dordrecht, 1987, http://dx.doi.org/10.1007/978-94-015-7744-1.
[16] V. Granville, M. Krivanek, J.-P. Rasson, Simulated annealing: a proof of convergence, IEEE Trans. Pattern Anal. Mach. Intell. 16 (1994) 652–656, http://dx.doi.org/10.1109/34.295910.
[17] Z.-H. Zhan, J. Zhang, Y. Li, Y.-H. Shi, Orthogonal learning particle swarm optimization, IEEE Trans. Evol. Comput. 15 (2011) 832–847, http://dx.doi.org/10.1109/TEVC.2010.2052054.
[18] M.E.H. Pedersen, A.J. Chipperfield, Simplifying particle swarm optimization, Appl. Soft Comput. 10 (2010) 618–628, http://dx.doi.org/10.1016/j.asoc.2009.08.029.
[19] F.R. Gonzalez, A.M. Nardillo, Retention index in temperature-programmed gas chromatography, J. Chromatogr. A 842 (1999) 29–49, http://dx.doi.org/10.1016/S0021-9673(99)00158-2.
[20] E.V. Dose, Simulation of gas chromatographic retention and peak width using thermodynamic retention indexes, Anal. Chem. 59 (1987) 2414–2419, http://dx.doi.org/10.1021/ac00146a020.
[21] C.A. Claumann, A. Wüst Zibetti, A. Bolzan, R.A.F. Machado, L.T. Pinto, Fast and accurate numerical method for predicting gas chromatography retention time, J. Chromatogr. A 1406 (2015) 258–265, http://dx.doi.org/10.1016/j.chroma.2015.06.004.
[22] S. Chapra, R. Canale, Numerical Methods for Engineers, 6th ed., McGraw-Hill Education, Boston, 2010, http://dx.doi.org/10.1016/0378-4754(91)90127-O.
[23] J. Randon, L. Maret, C. Ferronato, Gas chromatography-mass spectroscopy optimization by computer simulation, application to the analysis of 93 volatile organic compounds in workplace ambient air, Anal. Chim. Acta 812 (2014) 258–264, http://dx.doi.org/10.1016/j.aca.2014.01.016.
[24] T.M. McGinitie, H. Ebrahimi-Najafabadi, J.J. Harynuk, A standardized method for the calibration of thermodynamic data for the prediction of gas chromatographic retention times, J. Chromatogr. A 1330 (2014) 69–73, http://dx.doi.org/10.1016/j.chroma.2014.01.019.