On the inherent data fitting problems encountered in modeling retention behavior of analytes with dual retention mechanism

Eva Tyteca*, Gert Desmet

Department of Chemical Engineering (CHIS-IR), Vrije Universiteit Brussel, Pleinlaan 2, Brussels, Belgium

ARTICLE INFO

ABSTRACT

Some valuable insights have been obtained in the inherent fitting problems when trying to predict the retention time of complex, multi-modal retention modes such as encountered in HILIC and SFC. In this study, we used mathematical models with known input parameters to generate different sets of numerical test curves representative for systems exhibiting a complex, non-LSS dual retention behavior. Subsequently, we tried to fit these data sets using some popular (non-linear) literature models. Even in cases where a physical fitting model exists (e.g., the mixed model in case of pure additive adsorptive and partitioning retention), the fitting quality can only be expected to be relatively good (prediction errors expressed in terms of a normalized resolution error ε_{Rs}) when carefully selecting the scouting runs and the appropriate starting values for the fitting algorithm. The latter can best be done using a comprehensive grid search scanning a wide range of different starting values. This becomes even more important when no good physical model is available and one has to use a non-physical fitting model, such as the empirical Neue-model. The use of higher-order models is found to be quasi indispensable to keep the prediction errors on the order of some $\Delta R_s = 0.05$. Also, the choice of the scouting runs becomes even more important using these higher-order models. For highly retained compounds we recommend using scouting runs with long t_G/t_0-values or to include a run with a higher fraction of eluting solvent at the start of the gradient. When trying to predict gradient retention, errors with which the isocratic retention behavior is fitted are much less important for high retention factors k than errors made in the range of k near the one at the point of elution. The results obtained with a so-called segmented Neue-model (containing 7 parameters) were less good and thus practically not interesting (because of the high number of initial runs).

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

In reversed-phase liquid chromatography (RPLC), the use of retention models such as the linear solvent strength (LSS)-model [1] to predict the retention factor under gradient elution has become a great success and even lead to the introduction of several commercial software packages that allow the optimization of the gradient conditions via chromatogram simulation [2–8].

$$\ln(k) = \ln(k_w) - S\varphi \quad \text{[LSS model]} \tag{1}$$

where φ is the fraction of organic solvent. The same holds for normal phase (NPLC) and ion chromatography, where a similar simple isocratic retention model exists to accurately predict gradient retention factors (Eq. (2)).

$$\ln(k) = \ln(k_w) - S\ln(\varphi) \quad \text{[LOG model]} \tag{2}$$

where φ is the fraction of the strongest organic solvent. Eqs. (1) and (2) respectively describe a partitioning and an adsorptive retention mechanism.

In hydrophilic interaction chromatography (HILIC) and supercritical fluid chromatography (SFC) the situation becomes different. In HILIC the retention mechanism in these elution modes is generally accepted to be a combination of both partitioning and adsorption [9], although also electrostatic interactions can play an important role in the case of ionic compounds. SFC retention is considered multi-modal and highly dependent on the nature of the stationary phase [10–13], resulting in a much more pronounced non-linear $\ln(k)$ vs. ϕ-relationship than in RPLC. Two of the most

* Corresponding author. Tel.: +32 2693617.
E-mail address: eva.tyteca@vub.ac.be (E. Tyteca).

frequently used models proposed in literature to cope with this non-linear retention behavior are the following [14–19]:

$$\ln(k) = \ln(k_w) - S_1\varphi - S_2 \ \ln(\varphi) \quad \text{[Mixed model]} \tag{3}$$

$$\ln(k) = \ln(k_w) + 2 \cdot \ln(1 + S_2\varphi) - \frac{S_1\varphi}{1 + S_2\varphi} \quad \text{[Neue-model]} \tag{4}$$

where φ is the fraction of water (in the case of HILIC) or MeOH (in the case of SFC). The present study originated from the general observation that it appears to be much more difficult to accurately predict retention times in HILIC [18] and SFC [19] than in RPLC and NPLC. These difficulties already start in the model fitting stage, where sometimes severe convergence problems are encountered. Since convergence problems are difficult to evaluate in a real situation (for the actual solution is unknown) the present study has been set up to investigate the modeling ability of several theoretical dual retention systems, generating retention data by assuming the most simple of all possible dual retention relationships, i.e., by plainly adding two individual mechanisms via:

$$k = \omega \cdot k_1 + (1 - \omega) \cdot k_2 \tag{5}$$

where k_1 and k_2 respectively represent the retention factor of the first and the second retention mechanism, and α is a normalized weighting factor (see Table 1 for the k_1 and k_2 equations used in this study).

We will further refer to Eq. (5) as the numerical test curve. In each "experiment", this numerical test curve is first used to generate a limited series of typical isocratic or gradient scouting run retention time data. Subsequently, these scouting run data are fitted with any of the considered non-LSS parameter fitting models involving either three parameters (Eqs. (3) and (4)) or more (Eqs. (14)–(17) in Section 3.2.2.3 and Eqs. (18)–(21) in Section 3.3.3) to yield a set of best-fit retention parameters for the fitting model (e.g., k_w, S_1 and S_2 for the Neue-model in Eq. (4)). Finally, the fitting model was used to predict the retention time over the full range of possible isocratic and gradient elution conditions. Comparing the predicted retention times with those calculated using the numerical test curve, the fitting ability of the employed non-LSS fitting model can be directly quantified.

To select the fitting model, we mainly focused on models that are amenable to an analytical solution for the availability of such a solution greatly simplifies the fitting strategy. As a consequence, emphasis was in most cases put on the Neue-model. In an attempt to improve its fitting ability, higher order variants of the Neue-model have been established as well (cf. Eqs. (14)–(21)).

To vary the degree of complexity of the individual retention models underlying the investigated dual retention mechanisms, the k_1- and k_2-parameters in Eq. (5) have been calculated using 2-parameter as well as 3-parameter models. The experiences gained from these cases are respectively presented in Sections 3.2 and 3.3.

We would like to stress that in order to experimentally verify the retention modeling, not only a high number of isocratic runs is needed, but also a high number of gradient runs should be performed (to cover the full range of gradient conditions), as the prediction accuracy is highly dependent on the nature of the employed gradient profile. To conduct these experiments both in HILIC and in SFC, and, for poorly and highly retained compounds is practically almost not feasible. Moreover, although the data generation is performed in a pure mathematical way, the resulting isocratic retention curves were very similar to those obtained in previous studies [18,19]. As a justification for this statement, we included some experimental isocratic HILIC and SFC data (Supplementary Material, Fig. S-1), fitted with the employed mathematical model (Eq. (5)). We added a sum of two LSS-models, LSS- and LOG-model and two Neue-models. The latter gave the best fit of the isocratic data, indicating the complex retention behavior in

these chromatographic modes. The corresponding residuals are also given in Fig. S-1.

2. Numerical methods

2.1. Models for data generation

Instead of experimentally producing retention data, these were replaced by a series of theoretical isocratic and gradient retention data generated using Eq. (5) in combination with any of the models given by Eqs. (1)–(4) with known coefficients as the input for k_1 and k_2 (the different considered cases and their parameter values are given in Table 1). In the case of gradient elution, retention times were predicted by numerically integrating the fundamental gradient equation [20], based on the trapezoid rule:

$$t_0 = \int_0^{t_R - t_0} \frac{dt_s}{k(\varphi)} \tag{6}$$

Taking into account the instrument dwell time t_D (during which $k = k_0$) this equation becomes:

$$t_0 = \frac{t_D}{k_0} + \frac{1}{\beta} \cdot I \quad \text{with} \quad I = \int_{\varphi_0}^{\varphi_{elution}} \frac{d\varphi}{k(\varphi)} \tag{7}$$

where $\beta = \Delta\phi/t_G$. The percentage of ACN at elution $\varphi_{elution}$ is obtained by numerically searching the value for which $I - \beta \times (t_0 - t_D/k_0) = 0$. Inserting $\phi_{elution}$ in the retention models making up Eq. (5), the retention factor at the moment of elution ($k_{elution}$) is readily obtained. The effective gradient retention factor k_{eff} is calculated via:

$$k_{eff} = \frac{t_R - t_0}{t_0} = \frac{t_D}{t_0} + \frac{\phi_{elution} - \phi_0}{\beta t_0} \tag{8}$$

The data generation model was used at two instances. First a limited series of "scouting run" data was generated, typically including either five isocratic runs (with $\phi = 0.05$, 0.1, 0.15, 0.2 and 0.25) or 3 gradient runs (with composition running between $\phi = 0.05$ to 0.35 and $t_G = 5$, 10 and 15 min for a column with $t_0 = 1$ min) (=base case). To test the quality of the actual modeling step described in Section 2.2, the data generation model was also used to generate retention data for the full range of possible isocratic and gradient conditions (Section 2.3).

We would like to remark that for SFC prediction, pressure effects inside the column the retention factor k will not be constant everywhere inside the column, for every ϕ. Indeed, retention in SFC is the result of both the mobile phase composition ϕ and the pressure. These pressure effects might explain why better prediction accuracy was obtained using gradient scouting runs vs. isocratic scouting runs [19]. Extending Eq. (6) with the (up to date not yet fully understood) effect of pressure could result in better prediction accuracy in SFC, based on isocratic scouting runs.

2.2. Parameter fitting procedures

Without making any use of the knowledge of the models used to generate the "scouting run" data, these scouting run data were fitted with any of the considered non-LSS parameter fitting models involving either three (Eqs. (3) and (4)) or more parameters (Sections 3.2.2 and 3.3.3). We fitted k instead of $\ln(k)$, i.e., fitting $k = \exp(f(\varphi))$ instead of $\ln(k) = f(\varphi)$, because k is more relevant for prediction purposes than $\ln(k)$. Moreover, a *least-square* fitting in semi-log scale gives symmetric residuals in $\ln(k)$, resulting in severe absolute errors at high k-values, i.e. highest %ACN. Generally it is more interesting to have uniform predictions for medium to high k values because resolution needs retention.