ELSEVIER

Journal of Chromatography A, 842 (1999) 65–77

JOURNAL OF CHROMATOGRAPHY A

Application of statistical-overlap theory to gas chromatograms simulated on nonpolar stationary phases with commercial software

Clint Samuel, Joe M. Davis*

Department of Chemistry, Southern Illinois University at Carbondale, Carbondale, IL 62901-4409, USA

Abstract

Scores of gas chromatograms of complex nonpolar mixtures were simulated on nonpolar stationary phases with commercial software. The mixtures included petroleum, alkanes, pesticides, flavors and fragrances, drugs, volatiles, industrial solvents, phenoxy acids, semivolatiles, C-, N-, and S-containing polynuclear aromatic hydrocarbons, Arochlors, and polychlorinated biphenyls, naphthalenes, dibenzo-*p*-dioxins, and dibenzofurans. With commercial software, the retention times of mixture components were calculated from Kováts retention indices, and the standard deviations of single-component peaks were calculated from the modified Golay–Giddings equation. These peak parameters then were used to predict the numbers of peaks expected in the chromatograms, using laboratory software based on the point-process statistical theory of overlap and the Poisson distribution. In most cases, the predictions agreed very well with the numbers of peaks found in chromatograms simulated with the peak parameters. This agreement verified that the interval between successive single-component peaks in most gas chromatograms of complex nonpolar mixtures developed on nonpolar stationary phases can be represented by a random variable. Statistical-overlap theory can be applied to such chromatograms. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Petroleum; Retention indices; Peak overlap; Computer simulation

1. Introduction

Several theories have modelled by statistical means the overlap of peaks in one-dimensional separations of complex mixtures [1–6]. These theories are principally of two types. In the first, the intervals between adjacent peaks are interpreted using point-process statistics [1,2,5]; in the second, the frequency content of the separation is interpreted using Fourier analysis [3,4]. The basis of both types of theory is that the elution times of components in complex mixtures cannot be controlled completely by the researcher. Thus, the interval between adjacent peaks is a random variable that can be described using probability theory. The theories and their applications have been reviewed [7–9].

The point-process overlap theory associated with the senior author is now more sophisticated than that first proposed by Giddings and him [1]. The latter theory was extended to address systematic variations of the density [10] and standard deviation [10,11] of single-component peaks (i.e., peaks produced by pure mixture components) in separations, to predict mixture attributes in small regions of separations [12], and to describe overlap in separations using statistics other than Poisson statistics [13]. Recently, it also was extended to predict the total number of peaks [5,14] and numbers of singlet, doublet, triplet, etc., peaks [15] in separations having severe overlap. These latter extensions required a re-examination of

*Corresponding author.

0021-9673/99/$ – see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S0021-9673(98)00981-9

resolution and removed many ambiguities then existing in theory, which now rests on fairly solid physical principles.

In spite of this status, one might ask, ''So what?'' The question is legitimate. Extensive tests of point-process overlap theory by simple simulations of separations have verified its predictions [5,10,11,13–17]. The problem is that these simulations tested the mathematical aspects of theory but not the validity of the assumptions on which theory is based. For example, the interval between single-component peaks (i.e., SCPs) in these simulations was modelled by a random variable. Such modelling is a good test of any theory based on this assumption but it does not test the validity of the assumption itself.

In addition to simulations, point-process overlap theory has been applied to experimental chromatograms. These applications were of two types. The first, which are few in number, were to chromatograms containing known numbers of SCPs [18,19]. Here, the predictions of theory were verified. The latter, which are larger in number, were to poorly characterized chromatograms containing unknown numbers of SCPs [5,10,12,16,20–25]. These predictions were not confirmed; indeed, one purpose of overlap theory is to make predictions of attributes that cannot be measured easily. The researchers making the applications (including the senior author) assumed that the predictions were correct, provided that certain criteria were met. Not everyone shares this assumption, however, and perhaps one reason is that too few applications verifying theory have been reported.

One action that could improve confidence in overlap theory is its successful application to a large number of fully characterized mixtures differing widely in chemical composition. The simplest such mixture is a synthetic one prepared from pure standards; since its composition is known, the predictions of theory can be tested. The practical limitations of such a study, however, are large. First, many costly high-purity standards would be needed; secondly, one would have to arbitrarily select or reject standards for inclusion in mixtures; and finally, the time required for the chromatographic analyses would be large.

A realistic alternative is to simulate separations using software having sound bases in thermodynamics and transport properties (e.g., viscosity, diffusion coefficients, etc.). In particular, gas chromatography is mimicked with relative ease. In this paper, we develop simulations of gas chromatograms using the commercial software, Pro ezGC, and compare the numbers of peaks in these simulations to ones predicted by point-process overlap theory. Specifically, scores of temperature-programmed chromatograms of several mixtures of compounds, e.g., volatiles, industrial solvents, petroleum, pesticides, flavors and fragrances, polynuclear aromatic hydrocarbons containing C, N, and S atoms, Arochlors, and polychlorinated biphenyls, naphthalenes, dibenzo-*p*-dioxins, and dibenzofurans, were simulated on the stationary phases, 100% methylsilicone and 5% phenyl-/95% methylsilicone. The numbers of peaks so determined were compared to those predicted by theory. Close agreement was interpreted as a successful description of that gas chromatogram by theory.

We emphasize that these simulations differ markedly from those reported earlier by us and others, and their purpose also differs. In early simulations, the interval between SCPs was *forced* to be a random variable, and the simulations only confirmed mathematical aspects of theory. In the Pro ezGC simulations, however, the interval between SCPs is determined by thermodynamics and transport properties. In other words, *no statistical hypothesis is made*. Consequently, these simulations assist in testing if the interval between SCPs in complex gas chromatograms *is* a random variable, and the test is made using a theory whose rigor has been established when the interval is indeed a random variable.

A reasonable question to ask is, ''Why has such a study not been made until now?'' The answer is that, until recently, point-process overlap theory was not powerful enough to describe overlap in any but the most carefully controlled separations. This shortcoming no longer exists, as we shall show.

2. Theory

The theory below is a composite of several extensions, one of which is introduced here for the first time, of the simple result first published by Giddings and the senior author [1]

$$p = \bar{m} \exp(-\alpha) \quad (1)$$

where p, $\bar{m}$, and α are defined below. While insightful, Eq. (1) has too many limitations to be very useful (see below).

In the extended theory used here, each single-component peak (i.e., SCP) is represented by a point at its center. The SCPs (points) are distributed along separation coordinate x, which parallels the separation axis. The intervals between adjacent SCPs (i.e., adjacent points) are described by a probability density function (pdf), $h(x, z)$, where z is a coordinate that also parallels the separation axis and whose origin is an SCP center. The number, $h(x, z)\mathrm{d}z$, is the probability that the subsequent SCP lies between z and $z+\mathrm{d}z$.

If $h(x, z)$ is the inhomogeneous Poisson pdf [26]

$$h(x, z) = \lambda(x) \exp\left(-\int_0^z \lambda(u)\mathrm{d}u \right) \quad (2)$$

where $\lambda(x)$ is the density of SCPs (i.e., the average number of SCPs per unit time) at x and u is a dummy variable, then the average number p of peaks expected in a separation of extent X is [10]

$$p = \int_0^X \lambda(x) \exp[-\lambda(x)x_0(x)]\mathrm{d}x$$

$$= \bar{m} \int_0^1 f(\zeta) \exp[-f(\zeta)\alpha(\zeta)]\mathrm{d}\zeta \quad (3)$$

where $x=0$ is the beginning coordinate of the separation and $x_0(x)$ is the average minimum interval between adjacent SCPs required for separation at x. This interval is related to the average minimum resolution $R_S^*(x)$ required for separation at x by [1]

$$x_0(x) = 4\sigma(x)R_S^*(x) \quad (4)$$

where $\sigma(x)$ is the standard deviation of SCPs at x. By minimum resolution, one means that this resolution, or a larger one, results in separation. The reason that $R_S^*(x)$ is an average value is that different minimum resolutions are needed to separate two SCPs having different ratios of maximum heights (or intensities), and these different resolutions must be averaged [5,14,27,28].

In Eq. (3), reduced coordinate ζ equals x/X and $\bar{m}$ is the expected number of SCPs (i.e., points) in interval X. The pdf $f(\zeta)$ quantifies the density of SCPs at ζ, i.e., $f(\zeta)$ is large when many SCPs are found near ζ and small when they are not. The pdf $f(\zeta)$ is related to other aspects of Eqs. (2)–(4) by [10]

$$\lambda(\zeta) = \bar{m}f(\zeta)/X; \int_0^1 f(\zeta)\mathrm{d}\zeta = 1 \quad (5)$$

In Eq. (3), $\alpha(x)$ (or $\alpha(\zeta)$) is the saturation of the separation at x [1]

$$\alpha(x) = \lambda(x)x_0(x) = 4\lambda(x)\sigma(x)R_S^*(x);$$

$$\alpha(\zeta) = 4\bar{m}f(\zeta)\sigma(\zeta)R_S^*(\zeta) \quad (6)$$

where $\sigma(\zeta)=\sigma(x)/X$. The saturation is a measure of the amount of overlap at x, i.e., if $\alpha(x)$ is large, then overlap at x is high; if $\alpha(x)$ is small, then overlap at x is low.

Although not stated explicitly in the above equations, R_S^* varies with α (i.e., $R_S^*=R_S^*(\alpha)$) and this variation can be predicted by theory [5,14]. The variation occurs because resolution is defined for only two SCPs, whereas observed adjacent peaks in saturated separations often contain more than one SCP.

The new extension of theory reported here is the description of overlap in saturated separations having variable SCP density. Earlier extensions addressed separately the descriptions of overlap in unsaturated separations having variable SCP density [10] and saturated separations having constant SCP density [5,14]. It is simple to describe both simultaneously. The consequence is that R_S^*, which was constant in the former description and varied with α in the latter, now varies with both α and x since α itself varies with SCP density $\lambda(x)$ and standard deviation $\sigma(x)$. Rather than signify this variation by the awkward symbol, $R_S^*(x, \alpha(x))$, we will use $R_S^*(x)$. The extension has been confirmed by simulation [29].

Eqs. (2)–(6) clearly are more complicated than is Eq. (1) but they are the basis of a better overlap theory. The reasons are three-fold: (1) the density of SCPs varies throughout the separation, instead of being constant; (2) the standard deviations of SCPs vary throughout the separation, instead of being constant; and (3) theory predicts overlap in both well

resolved and poorly resolved separations, instead of only well resolved separations.

Equations that are similar to Eqs. (2)–(6) can be written for any inhomogeneous pdf. Only the Poisson pdf is investigated here, however, because it is the most appropriate pdf to use when a very large number of homologous series are present in a mixture [3,30]. In this paper, we principally are interested in how well point-process overlap theory based on the Poisson pdf describes overlap in realistic simulations of gas chromatograms of nonpolar compounds on nonpolar phases.

3. Procedures

3.1. Gas chromatographic simulations

The software, Pro ezGC for Windows, Version 2 (Restek, Bellefonte, PA, USA), was used with minor modification of its libraries. These libraries contain compounds identified as mixture components by various researchers and entered into the database by the software supplier. However, they are not comprehensive, i.e., unidentified compounds are not (and cannot be) included. Most libraries were used as distributed. We compiled additional libraries by grouping compounds in the Pro ezGC database according to various classifiers of interest to us, e.g., drug type, atom type, isotopic substitution, etc.

Several software-supplied libraries contained hundreds of compounds, whereas Pro ezGC simulates chromatograms having only 100 components or less. With such libraries, we mimicked ''windows'' of chromatograms containing as many contiguously eluting compounds as possible, with different windows containing different compounds. Most windows were chosen to span certain carbon-number ranges, but since we often used as many compounds as possible to obtain good statistics, windows sometimes contained a few compounds having either one more or one less carbon than indicated by the range.

The retention times and baseline widths of SCPs generated by compounds in Pro ezGC were calculated under the following conditions: mobile phase, helium; mobile phase delivery, constant pressure at an initial linear velocity of 30 cm/s; capillary length and I.D., 30 m and 250 μm; stationary-phase thickness, 0.25 μm. A single temperature program, 40°C to 250°C at 4°C/min with subsequent isothermal operation at 250°C, was used first, and the initial temperature then was raised, if necessary, to elute the first SCP within 10 min. On one occasion, a final temperature of 290°C was required. Successive windows of a complex chromatogram were developed using the same temperature program. Most separations were completed within 30 min; all were completed within 60 min. The stationary phases investigated were 100% methylsilicone (e.g., RTX-1, DB-1, etc.) and 5% phenyl-/95% methylsilicone (e.g., RTX-5, DB-5, etc.).

The SCP retention times were calculated by Pro ezGC from Kováts retention indices either reported in the literature or measured by the software distributer. The SCP widths at baseline were calculated by Pro ezGC using the modified Golay–Giddings equation and then increased 11.1% (the software default value) to mimic more realistic conditions (i.e., observed efficiencies usually are less than theoretical ones). These baseline widths then were converted to standard deviations by dividing by 4.

ASCII files of these retention times and standard deviations then were created and read by a FORTRAN program written in-house to simulate chromatograms. In the chromatograms, SCPs were Gaussians having the exponentially distributed amplitudes (or intensities) that closely mimic the amplitudes of SCPs in many natural-product mixtures [21,31,32].

Specifically, a chromatogram $c(t)$ containing m components was calculated at discrete times t as

$$c(t) = \sum_{i=1}^{m} A_i \exp[-(t - t_{r_i})^2/(2\sigma_i^2)] \quad (7)$$

where t_{r_i} and σ_i are the retention time and standard deviation of the ith SCP, as predicted by Pro ezGC, and A_i is an exponentially distributed random number. Chromatograms containing as few as five and as many as 100 components were calculated.

For each input file of t_{r_i}'s and σ_i's, 100 simulations were developed with different A_i's. This action mimicked a random change of component concentrations in the chromatogram but left unchanged the retention times and standard deviations. The numbers of peaks, as identified by maxima, in the 100 simulations were counted, and averages and standard deviations were calculated.

3.2. Prediction of p

The retention times and standard deviations calculated by Pro ezGC also were used to predict *p*, Eq. (3), by a FORTRAN program written in-house.

For each input file, the frequency $f(\zeta)$ was estimated from retention times by computing the discrete cumulative distribution function $F(\zeta)$ at equally spaced values of ζ and then numerically differentiating it, i.e., $f(\zeta)=\mathrm{d}F(\zeta)/\mathrm{d}\zeta$. The value of $F(\zeta)$ was computed as the fraction of reduced retention times $(t_{r_i}-t_{r_1})/X$ less than or equal to ζ, with X equal to $t_{r_m}-t_{r_1}$, the difference between the retention times of the first (t_{r_1}) and *m*th (t_{r_m}) component. The basis of these computations is explained elsewhere [10].

The number of coordinates used to compute $F(\zeta)$ and $f(\zeta)$ increased with *m*. In general, a large number of coordinates is desirable, because the variation of $f(\zeta)$ with ζ then is reasonably modelled. However, accurate values of $f(\zeta)$ can be calculated only if several components also have retention times near ζ; otherwise $f(\zeta)$ is undersampled. For $m>20$, the coordinate number was $\mathrm{INT}(m/10)+1$, where INT is the integer function (e.g., for $m=42$, this number was 5, with ζ equal to 0, 0.25, 0.50, 0.75, and 1); for $m\leq 20$, it was 3. Numerical derivatives were calculated using either quadratic or linear approximations (the latter was used at $f(\zeta)$'s boundaries). Frequency $f(\zeta)$ was represented in Eq. (3) by a 301-point linear interpolation of these derivatives. The interpolation was integrated between $\zeta=0$ and $\zeta=1$ to confirm the normalization constraint, Eq. (5).

An approximation to reduced standard deviation $\sigma(\zeta)$ was made by taking the σ_i values predicted by Pro ezGC, dividing them by $X=t_{r_m}-t_{r_1}$, and fitting them to a polynomial of degree 9 or less. The function $\sigma(\zeta)$ in Eq. (3) was represented by the polynomial.

The dimensionless integral for *p*, Eq. (3), was evaluated numerically using Simpson's rule, with $\alpha(\zeta)$ equal to Eq. (6). Because different R_S^*'s are associated with different α's (see below), Eq. (6) was evaluated iteratively by bisection at each node of the integration, with the functions $f(\zeta)$ and $\sigma(\zeta)$ described as detailed above and with $\bar{m}$ approximated by *m*, until the appropriate (α, R_S^*) coordinate was found. The evaluations were made using a linear interpolation of coordinates, (α, R_S^*), previously calculated for the Poisson pdf.

3.3. Comparison of simulation results and p

The average number of maxima in the gas-chromatographic simulations was compared to *p*. Close agreement was interpreted as support of the hypothesis that the interval between SCPs in that chromatogram is a random variable that can be described using point-process overlap theory based on Poisson statistics. Disagreement was interpreted as refutation of this hypothesis.

3.4. Miscellaneous

Pro ezGC was run on a 486 DX4 100 Mz personal computer operating under Windows 3.1. FORTRAN calculations were made using Language Systems FORTRAN (Language Systems, Sterling, VA, USA) on a Macintosh 4400 200 MHz Power personal computer and also FORTRAN PowerStation 32 (Microsoft, Redmond, WA, USA) on the 100 Mz personal computer. Graphs of, and fits to, σ_i/X were made using KaleidaGraph (Synergy Software, Reading, PA, USA) on either computer (versions 3.0.2 for Macintosh and 3.08 for 486 DXY).

4. Results and discussion

Fig. 1a is a graph of average minimum resolution R_S^* vs saturation α calculated from theory for the homogeneous Poisson pdf [5,14]. R_S^* decreases with increasing α, because adjacent observed peaks usually are composites of two or more SCPs at large α, and the two SCPs that actually are separated (one SCP in each observed peak) are fairly close together. The coordinates, (α, R_S^*), graphed as Fig. 1a are those at which Eq. (6) was evaluated iteratively at each node of the numerical integration of Eq. (3).

The insert in Fig. 1a is a graph of *p* vs α generated from simulations verifying that coordinates (α, R_S^*) do describe overlap in separations containing SCPs governed by Poisson pdfs. These simulations were *not* generated by Pro ezGC; rather, they contained 1000 SCPs whose retention times were

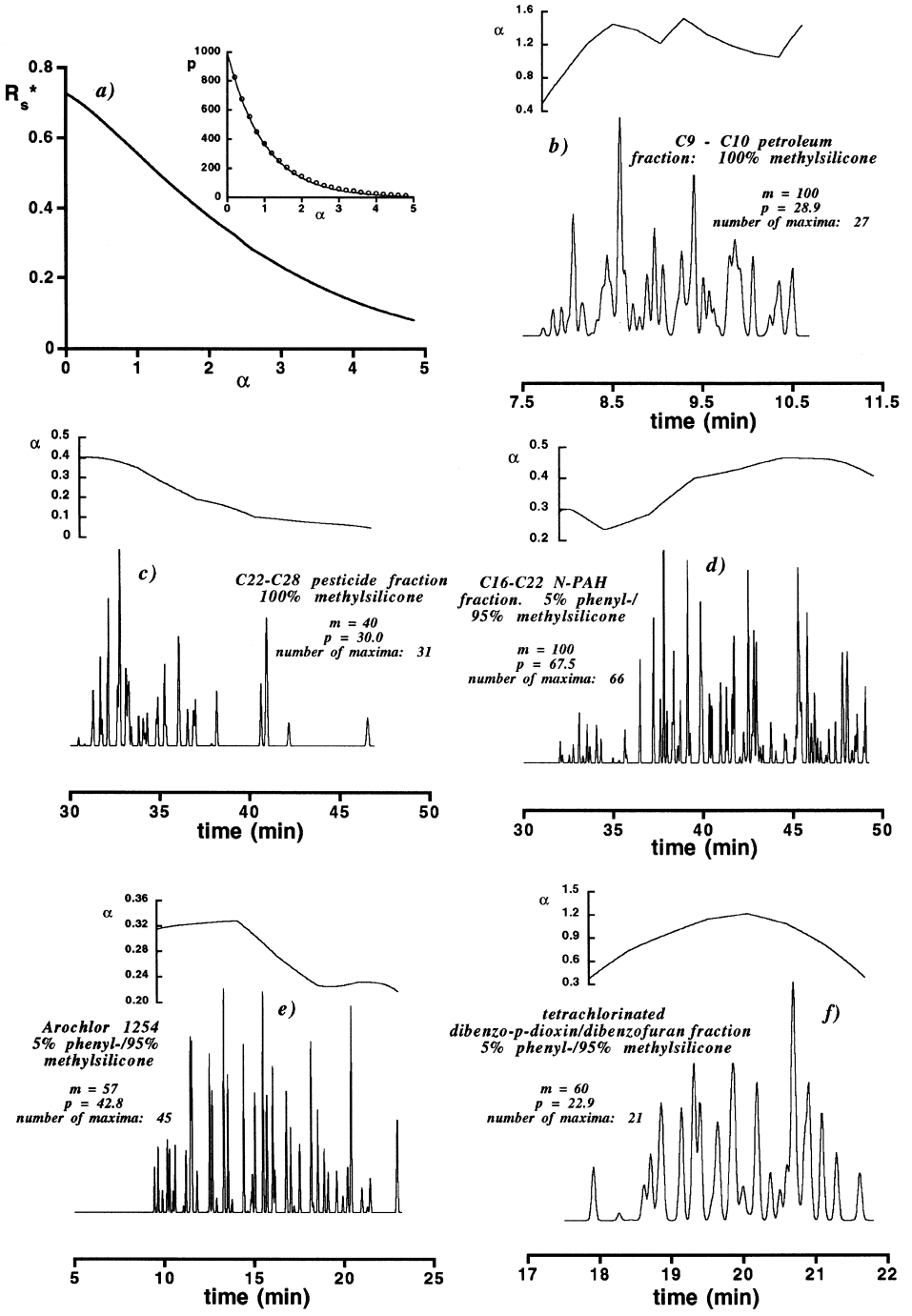

Fig. 1. (a) Graph of average minimum resolution R_S^* vs saturation α for the homogeneous Poisson pdf. Insert is graph of the average number p of peaks vs α calculated from simulations of separations governed by this pdf. (b–f) Simulated chromatograms (or fractions of chromatograms), as developed on nonpolar stationary phases, of (b) petroleum, (c) pesticides, (d) nitrogen-containing polynuclear aromatic hydrocarbons (N-PAHs), (e) Arochlor 1254, and (f) tetrachlorinated dibenzo-*p*-dioxins and dibenzofurans. Plots immediately above chromatograms are graphs of α vs time. Stationary phases, numbers m of components, p of expected peaks, and of maxima are reported.

forced to obey a homogeneous Poisson pdf and whose standard deviations were constant. This constant was calculated from Eq. (6) and the coordinates, (α, R_S^*). The symbols in the insert are the average numbers of maxima in 100 such simulations; the curve is the expected result, Eq. (1). The two agree closely, although a small disagreement is observed at large α. Thus, the coordinates (α, R_S^*) in Fig. 1a describe overlap well, if the intervals between successive SCPs obey Poisson statistics.

We turn to the Pro ezGC simulations to see if this behavior is observed. Fig. 1b–f are typical simulated gas chromatograms calculated from the retention times and standard deviations predicted by Pro ezGC. It is clear that overlap is severe in some cases, e.g., in Fig. 1b, only 27 maxima are obtained from 100 petroleum compounds spanning the C_9 and C_{10} *n*-alkanes. Inspection shows that both the density and widths of SCPs vary in the simulations. Immediately above the figures are graphs of saturation α vs time calculated from the predicted retention times and standard deviations. As expected, α is low when the separation is uncrowded and high when it is crowded. The discontinuities in α are caused by the discrete nature of the numerically determined frequency, $f(\zeta)$. It is clear that α varies in the simulations. Thus Eq. (1), which is based on constant α, is too simple to model overlap in realistic simulations of gas chromatograms. However, Eq. (3) does model overlap well. In the figures, the number *m* of components, the expected number *p* of peaks predicted by Eq. (3), and the number of maxima in the simulations are reported, and the latter two numbers agree closely.

Subsequent graphs depict on their vertical axes the numbers of peak maxima in simulations, as represented by open circles with error bars equal to one standard deviation, values of *p* predicted by Eq. (3), as represented by open squares, and values of *m*, as represented by open triangles. The horizontal axes represent various attributes, e.g., types of mixtures, fractions of mixtures, sources of mixtures, compounds targeted by US Environmental Protection Agency (EPA) methods, chemical functionality, etc. The abbreviations used to label these axes are defined in Table 1 (e.g., ''antihist'' stands for antihistamines). This type of axis is awkward, but we think such graphs are the simplest means to convey our results (tables would be horrific!). With figures, one can tell at a glance how simulation results, *p*, and *m* compare.

Our results should be viewed as a survey of numerous applications of overlap theory to nonpolar mixtures separable by gas chromatography, so each application is discussed only briefly. Fig. 2a–f depicts results for the stationary phase, 100% methylsilicone (e.g., RTX-1, DB-1, etc.), and mixtures comprised of petroleum, alkanes, pesticides, flavors and fragrances, drugs, volatiles, and industrial solvents. Fig. 2a shows results for petroleum compounds approximately spanning the C_1 to C_{21} *n*-alkanes, with the attribute equal to carbon-number range (e.g., results are reported for compounds roughly spanning the C_1–C_5 *n*-alkanes, the C_5–C_6 *n*-alkanes, etc.). In effect, the attribute represents different windows in a single complex petroleum chromatogram developed with a single temperature program. The agreement between simulation and theory is excellent, even when the number of peaks is only 20% of the number of components (e.g., as for the C_6–C_7 range). Because petroleum is a very complex mixture containing many homologous series, it satisfies conditions for the Poisson pdf and the agreement between simulation and theory is unsurprising. These results justify the use of the Poisson pdf by us [24] and others [16] in earlier descriptions of overlap in petroleum gas chromatograms. In contrast, Dondi et al. found that overlap in small domains of naptha gas chromatograms was described best by pdfs other than Poisson [33], and Felinger justified this finding using variance–time curves [30].

Fig. 2b actually is a composite of two unrelated figures and is presented as such for convenience. The left-hand side of Fig. 2b shows results for 32 of the C_1–C_{40} *n*-alkanes. Simulation and theory agree well and both show that separation almost is complete. The right-hand side of the figure shows results for pesticides, with an attribute equal to carbon-number range. As before, the attribute represents different windows in a single complex chromatogram. The agreement between simulation and theory is very good and a fair amount of overlap results, especially in the C_{18}–C_{22} range. In the different ranges, the number of peaks is only 56–75% of the number of components.

Table 1
Explanation of abbreviations of attributes in Figs. 2 and 3.

C*m*– C*n*	Fraction of chromatogram eluting between *n*-alkanes having carbon numbers *m* and *n*
pep/spm	Peppermint/spearmint/mint/eucalyptus oils
lem/lim	Lemon/lime/orange/mandarin oils
thym/ros	Thyme/rosemary/sage oils
pin ndle	Pine needle/dwarf pine oils
cnm/bois	Cinnamon/bois de rose oils
cara/dill	Caraway/dill seed oils
star anis	Star anise/fennel oils
bendiaz	Benzodiazepines
barbitur	Barbiturates
phnthiaz	Phenothiazines
antidepr	Antidepressants
anesth	Anesthetics
antihist	Antihistamines
CNS stm	Central-nervous system stimulants
antipsy	Antipsychotics
all cmpds	All relevant compounds in the Pro ezGC database
CLP	Chlorine-containing pesticides (internal library)
chlorine	Chlorine-containing pesticides (home-built library)
nitrogen	Nitrogen-containing pesticides
org-phos	Organophosphorus pesticides
n-*m* (*n*,*m* are numbers)	Fraction of PCB chromatogram whose first peak is the *n*th eluting component and last peak is *m*th eluting component
PCNs	Polychlorinated naphthalenes
sub2378	Chlorine-substituted 2,3,7,8 tetrachloro-dibenzo-*p*-dioxins/-dibenzofurans
m-*n* D/Fs	Dibenzo-*p*-dioxins/dibenzofurans substituted by anywhere from *m* to *n* chlorines

Note: All attributes that are numbers not preceded by the letter ''*A*'' and not containing dashes designate compounds targeted by the EPA method having that number; an ''*Rx*'' following said number designates revision *x*. All numbers preceded by the letter ''*A*'' designate Arochlors having that number.

Fig. 2c depicts results for flavors and fragrances, with an attribute equal to mixture source (e.g., pine needle). As before, theory describes overlap in both simple mixtures (e.g., components of peppermint and spearmint oils) and more complex ones (e.g., Sets 1 to 3). Sets 1–3 on the figure's right-hand side simply represent different windows in one complex chromatogram. For simple mixtures having less than 40 components, separation is almost complete as is confirmed by theory. In the complex mixtures, however, some overlap occurs and the number of peaks is 83–89% of the number of components. These findings justify this group's earlier predictions of overlap in complex gas chromatograms of lime and peppermint oils [24].

Fig. 2d depicts results for pharmaceutical drugs, with an attribute equal to drug family (e.g., barbiturates, opiates, antihistamines, etc.). The results on the figure's left-hand side were obtained using databases supplied by the software distributer; the other results were obtained using databases constructed in-house from compounds in Pro ezGC. In general, simulation and theory agree well, except for the more complex mixture of antidepressants. Both simulation and theory show that separation almost is complete, but this finding is unsurprising since the mixtures contain about 30 or fewer components.

Fig. 2e depicts results for volatiles, with an attribute equal to compounds targeted by EPA methods of analysis. However, the chromatographic conditions actually mimicked here differ in most cases from those mandated by EPA. Again, theory correctly describes the extent of overlap in the simulated chromatograms. The very simple mixture targeted by

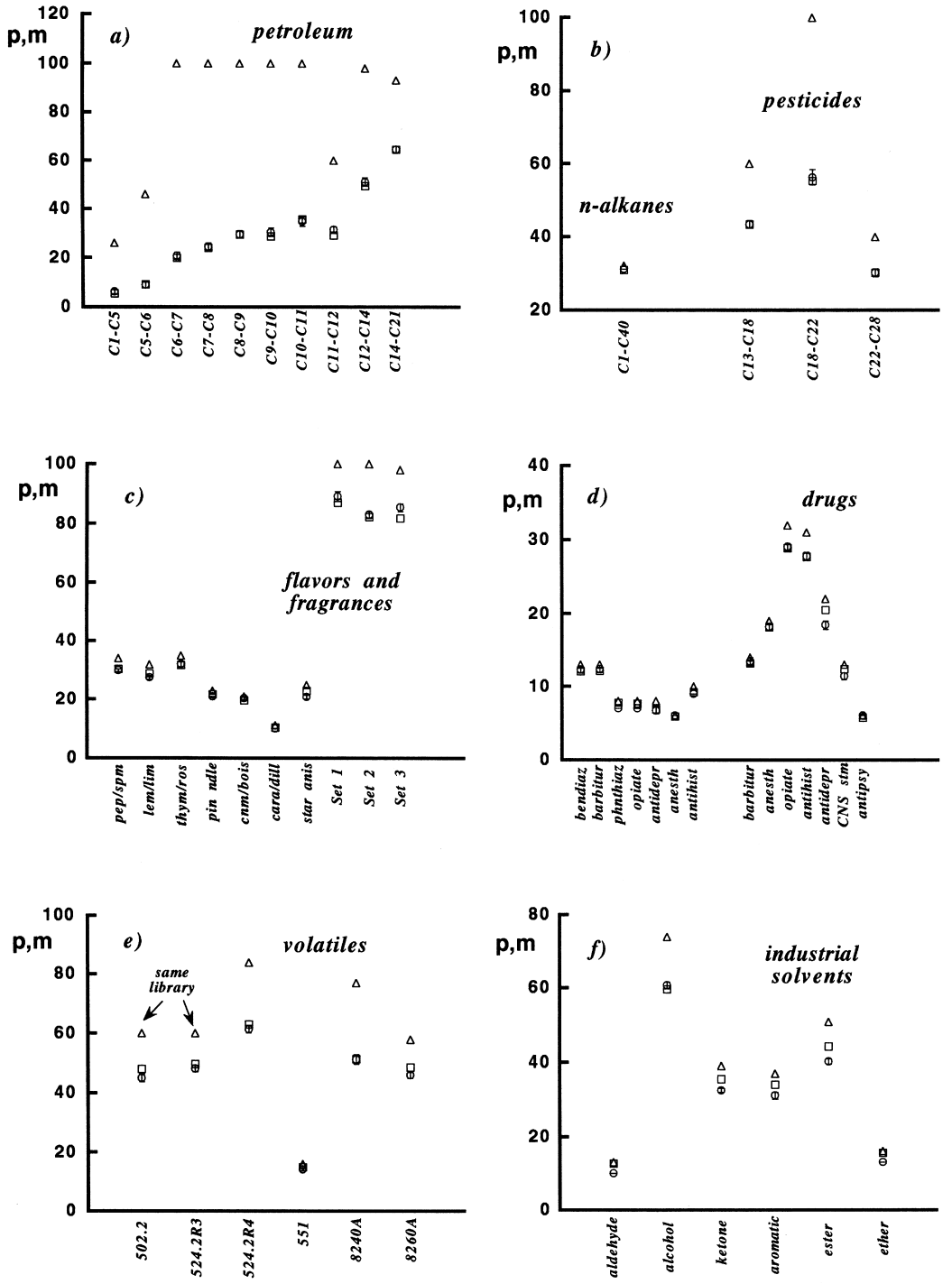

Fig. 2. Average numbers of peaks (○) and numbers *m* of components (△) in 100 chromatograms simulated on the stationary phase, 100% methylsilicone, and average numbers *p* predicted by theory (□), vs various attributes. Error bars are one standard deviation of peak numbers in simulations; attributes are described in text and in Table 1. (a) Petroleum, (b) *n*-alkanes and pesticides, (c) flavors and fragrances, (d) drugs, (e) volatiles, and (f) industrial solvents.

method 551 is almost completely separated. However, the more complex mixtures produce chromatograms containing overlapping peaks. For all methods, the number of observed peaks varies from 67 to 88% of the number of components.

Fig. 2f depicts results for industrial solvents, with an attribute equal to chemical functionality (e.g., alcohol, diethyl ether, etc.). Here, simulation and theory are close but do differ, and the disagreement often is significant (i.e., the theoretical p lies outside the bounds of the standard deviation of the number of simulated peaks). In general, theory predicts slightly more peaks than are observed.

Fig. 3a–f are graphs similar to those discussed, but for the stationary phase, 5% phenyl-/95% methylsilicone (e.g., RTX-5, DB-5, etc.). The mixtures investigated are petroleum, phenoxy acids, pesticides, semivolatiles, C-, N-, and S-PAHs (polynuclear aromatic hydrocarbons), Arochlors, polychlorinated biphenyls, polychlorinated naphthalenes, polychlorinated dibenzo-*p*-dioxins and dibenzofurans, and drugs. Fig. 3a is a composite of two unrelated figures (presented as such for convenience), with the results for petroleum on the left-hand side and for phenoxy acids on the right-hand side. The attribute of petroleum is carbon-number range, and results are presented as different windows in a single complex chromatogram. The agreement between simulation and theory is generally very good. Of some surprise is the slight disagreement between simulation and theory for the C_{13}–C_{18} range, since excellent agreement was found for nearly the same range on 100% methylsilicone (see Fig. 2a).

The right-hand side of Fig. 3a depicts results for phenoxy acids. Two of the three attributes identify compounds targeted by EPA methods; the third identifies all phenoxy acids in the database. The simple mixtures targeted by EPA methods are well separated and theory confirms this. However, a marked disagreement exists between theory and simulation, when all phenoxy acids are considered. As was the case for industrial solvents on 100% methylsilicone (see Fig. 2f), theory predicts a greater number of peaks than is found.

Fig. 3b depicts results for pesticides. The attributes on the figure's left-hand side are compounds targeted by various EPA methods. A second attribute on the figure's right-hand side is the active atom in the pesticide, e.g., chlorine, nitrogen, or phosphorus; the libraries of these compounds were built in-house. The compounds targeted by most EPA methods are well separated and theory confirms this finding. The separation of compounds targeted by method 1618 is incomplete, however, as is the separation of the large set of chlorinated pesticides. It is not surprising that these mixtures are not resolved fully, since they contain large numbers of compounds. In these two cases, theory also agrees well with simulation.

Fig. 3c depicts results for semivolatiles and PAHs containing C, N, and S atoms. The figure is comprised of three parts, with one part dedicated to each atom type. The attribute is carbon-number range and represents different windows in three different complex chromatograms (i.e., one for C-PAHs, one for N-PAHs, and one for S-PAHs). Overlap is extensive, and in general the numbers of peaks determined by simulation and predicted by theory agree well (although some small differences do exist). The exception is for S-PAHs spanning the range C_{11}–C_{25}; as in the disagreements reported earlier, theory predicts a greater number of peaks than is observed.

Fig. 3d depicts results for Arochlors, polychlorinated biphenyls (PCBs), and polychlorinated naphthalenes (PCNs) in a graph comprised of three parts. The attribute of the Arochlors is identification number (e.g., 1254). All 209 PCBs are considered, and the first 70 PCBs that elute are identified by ''1–70'', the next set of 70 by ''71–140'', and the final 69 by ''141–209''. In other words, the PCB attribute represents different windows in one complex chromatogram. The PCNs are treated separately, with the label ''PCN'' as the attribute.

The results show that overlap is fairly high in the different Arochlors, except for the simple 1221 mixture, and theory also predicts the amount of overlap. Similar results are found for the PCBs. Here, the overlap is especially severe, with the number of peaks equalling only 45 to 67% of the number of components. Again, theory correctly predicts these numbers. This finding is of particular interest, since overlap studies based on Fourier analysis show that these mixture types actually are better described by pdfs other than Poisson [34]. Simulation and theory diverge for the PCNs, however, and theory overestimates the number of peaks.

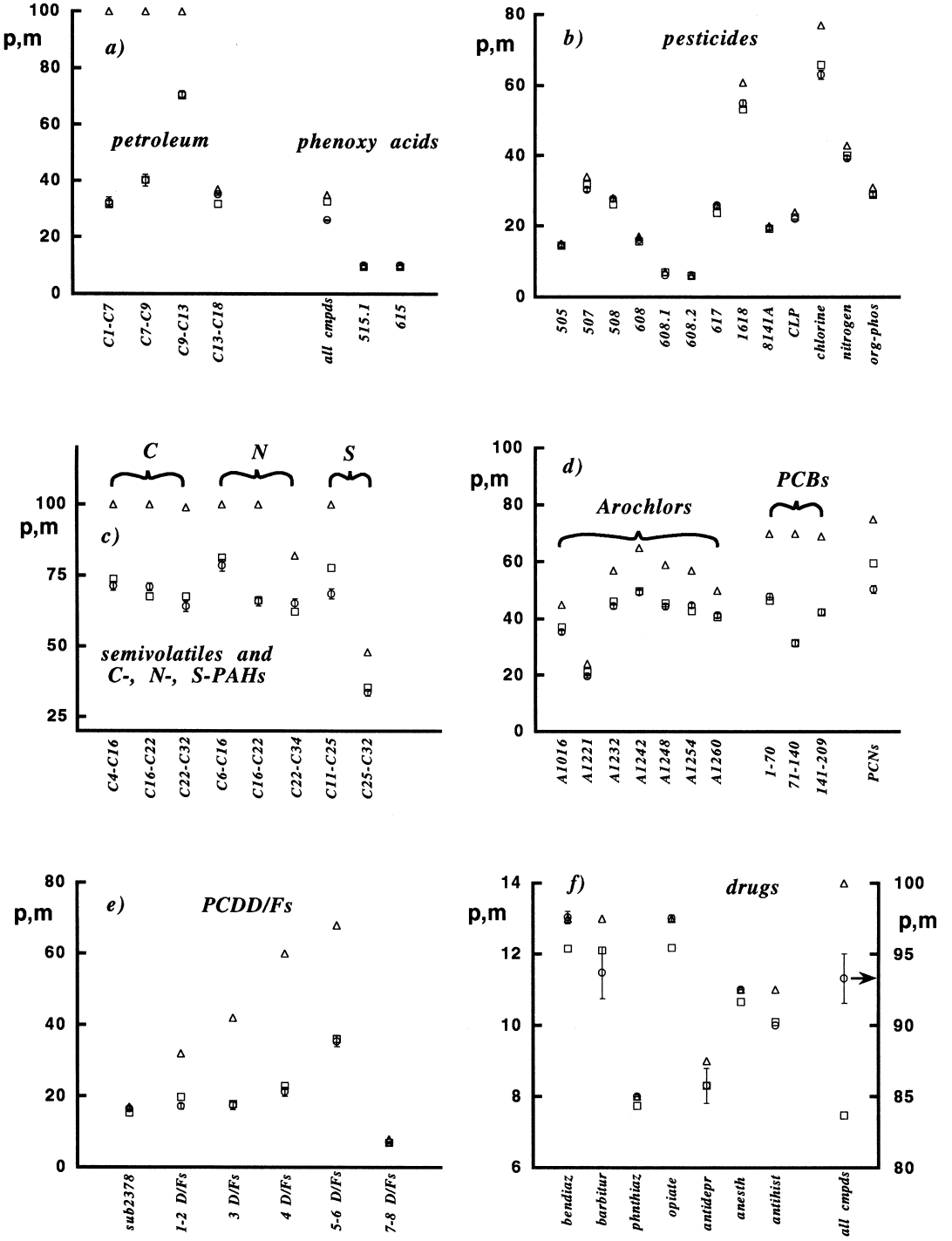

Fig. 3. As in Fig. 2, but the stationary phase is 5% phenyl-/95% methylsilicone. (a) Petroleum and phenoxy acids, (b) pesticides, (c) semivolatiles and C-, N-, and S-containing polynuclear aromatic hydrocarbons (PAHs), (d) Arochlors, polychlorinated biphenyls (PCBs), and polychlorinated naphthalenes (PCNs), (e) polychlorinated dibenzo-*p*-dioxins and dibenzofurans (PCDD/Fs), and (f) drugs.

Fig. 3e depicts results for polychlorinated dibenzo-*p*-dioxins (PCDDs) and dibenzofurans (PCDFs), or PCDD/Fs. The attribute, ''sub2378'', designates subsequent chlorine substitutions of the toxic isomers, 2,3,7,8 PCDD/F; the attribute ''*m*-*n* D/Fs'' designates PCDD/Fs substituted by anywhere from *m* to *n* chlorines. This second attribute represents different windows in a single complex chromatogram. The substituted 2,3,7,8 isomers are well resolved and theory confirms the separation. However, the substituted PCDD/Fs generally are subject to extensive overlap, as confirmed by theory.

Fig. 3f depicts results for pharmaceutical drugs. Two axes for *p* and *m* are shown, with the left-hand axis relevant to all results except those on the extreme right-hand side. In general, the simple mixtures containing 13 or fewer compounds are well resolved, and theory does a fair job in predicting this. However, the entire database of drugs is resolved incompletely, and theory underestimates the peak number. The underestimation is intriguing, since previous disagreements resulted from overestimating *p*.

5. Conclusions

It is tempting to suggest our findings reflect the behavior of real-world nonpolar mixtures subject to gas chromatography on nonpolar stationary phases. In truth, however, they actually reflect the behavior of mixtures comprised of compounds in the Pro ezGC database. The criteria for the inclusion of compounds in the database were the availability of literature data and the compounds' relevance to current legal, environmental, and health issues. Such a database, while extensive, does not necessarily represent real-world mixtures. However, it may well represent real-world mixtures in which chemists currently have interest.

With this thought in mind, we think our findings strongly vindicate the predictive power of statistical-overlap theory, and in particular the point-process overlap theory based on the Poisson pdf, as applied to complex gas chromatograms of nonpolar mixtures developed on nonpolar stationary phases. While this vindication is satisfying, criteria currently are lacking by which we can identify the low percentage of cases that do not follow theory. As shown by Felinger, variance–time curves are useful in this identification when the SCP density is constant [30]. In our simulations, the SCP density usually varied, and we do not know how to interpret variance–time curves calculated from equations derived for the constant-density case (the variance so determined is larger than that expected for the constant-density Poisson case). All we can state now is that theory *will* fail in a small number of cases and that we must find criteria to identify these cases.

In those cases to which theory does apply, a simple pattern is evident. Most mixtures containing roughly 40 compounds or less can be resolved or nearly resolved. In contrast, mixtures containing larger numbers of compounds are subject to overlap, and the amount of overlap increases with increasing numbers of compounds. The number, 40, is not talismanic, however, and is subject to variation as separation conditions change.

In closing, it is useful to consider extensions and improvements of our work. It easily can be expanded to examine other chromatographic conditions, e.g., different rates of temperature programming, column lengths and diameters, mobile phases, mobile-phase velocities, stationary-phase thicknesses, etc. Here, we used only typical separation conditions, with our criterion being completion of separation within a reasonable time. Undoubtedly, separation could be improved in some cases by use of different conditions (e.g., the separation of C_1–C_5 compounds in petroleum mixtures could be improved by reducing the initial capillary temperature). Of particular interest is the examination of polar stationary phases, e.g., polyethylene glycol. Furthermore, it would be most unwise to draw far-reaching conclusions from the overlap of compounds targeted by various EPA methods, since the actual analytical methods mandated by EPA were not used here. Finally, our chromatogram simulations could be improved by using algorithms that more realistically mimic SCP amplitudes than do exponential random-number generators (e.g., *n*-alkanes consistently produce strong responses in petroleum because of high abundance).

Acknowledgements

The authors thank John Garrett of Analytical

Innovations, Inc., for many helpful discussions of Pro ezGC. This work was supported by the National Science Foundation (CHE-9714328).

References

[1] J.M. Davis, J.C. Giddings, Anal. Chem. 55 (1983) 418.
[2] M. Martin, D.P. Herman, G. Guiochon, Anal. Chem. 58 (1986) 2200.
[3] A. Felinger, L. Pasti, F. Dondi, Anal. Chem. 62 (1990) 1846.
[4] A. Felinger, L. Pasti, F. Dondi, Anal. Chem. 63 (1991) 2627.
[5] J.M. Davis, Anal. Chem. 69 (1997) 3796.
[6] F. Dondi, A. Bassi, A. Cavazzini, M.C. Pietrogrande, Anal. Chem. 70 (1998) 766.
[7] J.M. Davis, Adv. Chromatogr. 34 (1994) 109.
[8] A. Felinger, Data Analysis and Signal Processing in Chromatography (Data Handling in Science and Technology, Vol. 21), Elsevier, 1998, Ch. 15 and 16, pp. 331; 385.
[9] A. Felinger, Adv. Chromatogr. 39 (1998) 201.
[10] J.M. Davis, Anal. Chem. 66 (1994) 735.
[11] J.M. Davis, J. Microcol. Sep. 9 (1997) 193.
[12] J.M. Davis, J. Microcol. Sep. 7 (1995) 3.
[13] M.C. Pietrogrande, F. Dondi, A. Felinger, J.M. Davis, J. Chemometr. Intell. Lab. Syst. 28 (1995) 239.
[14] J.M. Davis, Chromatographia 44 (1997) 81.
[15] J.M. Davis, J. Chromatogr. A 837 (1999) 187.
[16] D.P. Herman, M.-F. Gonnord, G. Guiochon, Anal. Chem. 56 (1984) 995.
[17] J.M. Davis, J.C. Giddings, Anal. Chem. 57 (1985) 2168.
[18] J.M. Davis, J. Chromatogr. 449 (1988) 41.
[19] S.L. Delinger, J.M. Davis, Anal. Chem. 62 (1990) 436.
[20] J.M. Davis, J.C. Giddings, Anal. Chem. 57 (1985) 2178.
[21] F. Dondi, Y.D. Kahie, G. Lodi, M. Remelli, P. Reschiglian, C. Bighi, Anal. Chem. Acta 191 (1986) 261.
[22] S. Coppi, A. Betti, F. Dondi, Anal. Chim. Acta 212 (1988) 165.
[23] F. Dondi, T. Gianferrara, P. Reschiglian, M.C. Pietrogrande, C. Ebert, P. Linda, J. Chromatogr. 485 (1990) 631.
[24] F.J. Oros, J.M. Davis, J. Chromatogr. 550 (1991) 135.
[25] M. Johll, D. Bowlin, J.M. Davis, Anal. Chim. Acta 302 (1995) 133.
[26] D.R. Cox, V. Isham, Point Processes, Chapman and Hall, New York, 1980.
[27] M.Z. El Fallah, M. Martin, Chromatographia 24 (1987) 115.
[28] A. Felinger, Anal. Chem. 69 (1997) 2976.
[29] J.M. Davis, unpublished results.
[30] A. Felinger, Anal. Chem. 67 (1995) 2078.
[31] L.J. Nagels, W.L. Creten, P.M. Vanpeperstraete, Anal. Chem. 55 (1983) 216.
[32] L.J. Nagels, W.L. Creten, Anal. Chem. 57 (1985) 2706.
[33] F. Dondi, A. Betti, L. Pasti, M.C. Pietrogrande, A. Felinger, Anal. Chem. 65 (1993) 2209.
[34] M.C. Pietrogrande, L. Pasti, F. Dondi, M.H.B. Rodriguez, M.A.C. Diaz, J. High Resolut. Chromatogr. 17 (1994) 839.