Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Chromatography A, 1096 (2005) 187–198
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Selection of orthogonal reversed-phase HPLC systems by univariate and auto-associative multivariate regression trees

R. Put [a], E. Van Gyseghem [a], D. Coomans [b], Y. Vander Heyden [a,*]

[a] *Department of Pharmaceutical and Biomedical Analysis, Pharmaceutical Institute, Vrije Universiteit Brussel-VUB, Laarbeeklaan 103, B-1090 Brussels, Belgium*
[b] *Statistics & Intelligent Data Analysis Group, James Cook University, Townsville 4814, Qld, Australia*

Available online 23 May 2005

Abstract

In order to select chromatographic starting conditions to be optimized during further method development of the separation of a given mixture, so-called generic orthogonal chromatographic systems could be explored in parallel. In this paper the use of univariate and multivariate regression trees (MRT) was studied to define the most orthogonal subset from a given set of chromatographic systems. Two data sets were considered, which contain the retention data of 68 structurally diversive drugs on sets of 32 and 38 chromatographic systems, respectively. For both the univariate and multivariate approaches no other data but the measured retention factors are needed to build the decision trees. Since multivariate regression trees are used in an unsupervised way, they are called auto-associative multivariate regression trees (AAMRT). For all decision trees used, a variable importance list of the predictor variables can be derived. It was concluded that based on these ranked lists, both for univariate and multivariate regression trees, a selection of the most orthogonal systems from a given set of systems can be obtained in a user-friendly and fast way.

Keywords: Orthogonal chromatographic systems; CART; Univariate regression trees; Multivariate regression trees; Method development; Unsupervised

1. Introduction

Reversed-phase high performance liquid chromatography (RPLC) is one of the most popular separation techniques used in the pharmaceutical industry [1]. Nowadays a wide range of columns containing different stationary phases are available and can be combined with a range of mobile phases. This diversity opens the opportunity to use RPLC for almost any pharmaceutical separation [2,3].

However, it is often not easy to select an appropriate starting point, e.g. a first choice of stationary and mobile phase, which is then optimized during further method development. Most of the time such a starting point is selected using trial-and-error methods [3]. An interesting approach is a setup in which different chromatographic systems are tested in parallel [4]. From these exploring experiments, the best initial starting conditions for a given new mixture are retained and they are further studied to improve the separation during further method development. In this context it would be ideal to have a set of chromatographic systems with different selectivities. In order to obtain such an ideal subset of chromatographic systems attempts are made to select the most orthogonal from larger sets [5–9]. Orthogonal chromatographic systems are defined as systems with a strongly different selectivity. This is caused by the fact that different mechanisms of retention are present in these systems or the retention is influenced by different charges of the given solutes [5,6].

Moreover, such an approach may be useful to reveal the composition of an unknown mixture. One of the most common mixtures to be separated in the pharmaceutical industry consists of a new drug and its impurities, which initially are unknown in number and structure. Exploring the separation on a set of parallel orthogonal systems can maximize the probability that all substances will be revealed and indicate

* Corresponding author. Tel.: +32 2 477 47 34; fax: +32 2 477 47 35.
E-mail address: yvanvdh@vub.ac.be (Y. Vander Heyden).

0021-9673/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2005.03.138

which system(s) are most appropriate for further method optimization.

The aim of this paper is to introduce and evaluate two new methodologies based on univariate [10] and multivariate regression trees [11] for the selection of orthogonal chromatographic systems. Univariate regression trees were proposed by Breiman et al. in 1984 [10] as one part of the non-parametric statistical method called classification and regression trees (CART). The name indicates that the methodology can handle both categorical and numerical variables as a univariate response [10]. CART is frequently used for modelling in different fields such as analytical chemistry, medical diagnosis, clinical epidemiology and ecology [12–15], since it can handle large sets of data (thanks to automatic feature selection), but probably more important, since the resulting models are simple decision trees, which are very easy to interpret. Multivariate regression trees (MRT) were introduced in 1992 by Segal to handle longitudinal data [16]. The methodology implemented in our study, was proposed by De'Ath [11] in the field of ecology. Another recently published paper by Larsen and Speckman [17] proposes an analogue method for the analysis of abundance data in ecology. In general, MRT's are proposed for the simultaneous description of several responses by tree models, using a set of independent variables to extract and predict the clusters present in the multivariate responses. More recently, Questier et al. [18] proposed to use MRT in an unsupervised way. Unsupervised means that no response variables are available in the data set, which means that the original variables are used not only as explanatory variables, but also as response variables. This approach was called auto-associative multivariate regression trees (AAMRT) and was proposed as a data mining cluster analysis method [18].

In order to study the selection of orthogonal systems, several chromatographic systems with different stationary phases, buffer pH, temperature and organic modifiers were examined. Two data sets were studied. The first consists of a set of eight silica-based systems, combined with four different pH levels in order to define a total of 32 chromatographic systems. The second consists of 38 chromatographic systems including 12 diverse columns at different mobile phase conditions. The selected subsets of orthogonal chromatographic systems are compared to those obtained using other methodologies on the same data sets [5,6,19]. These selections were based on Pearson's correlation coefficient color maps, in which the chromatographic systems were ranked according to increasing dissimilarities in a weighted pair group method using arithmetic averages (WPGMA)-dendrogram. In the literature, the orthogonal chromatographic systems usually are selected based on the evaluation of the Pearson's correlation coefficients [7,8] or parameters derived from it [9,19]. In order to enhance the interpretation of large correlation coefficients matrices for large sets of chromatographic systems, and to select the most orthogonal or the similar systems, several visualization methods were investigated, such as dendrograms from the hierarchical weighted-average-linkage clustering technique or weighted pair group method using arithmetic averages (WPGMA) method, OPTICS color maps and PCA [5,20]. Another approach suggested by Forlay-Frick et al. [19] is based on the generalized pairwise correlation method (GPCM) [21] for which the authors evaluated different statistical selection criteria. In this paper we evaluate two new methodologies for the selection of orthogonal chromatographic systems based on univariate and multivariate regression trees, and their use to sort the systems in color maps.

2. Theory

2.1. Univariate regression trees

2.1.1. Classification and regression trees

In 1984, Breiman et al. [10] introduced a statistical method for classification and modelling, called "classification and regression tree (CART) analysis". In this approach, a binary partitioning procedure is applied in order to explain the variation of a single dependent variable (the response variable or response) based on a set of independent variables (the explanatory variables or predictors). CART can handle both categorical and numerical variables as response and predictors. Since CART can handle only one response variable at the time, it is a univariate method. In this paper the term "univariate regression trees" is used, in order to emphasize the difference between this univariate method and the multivariate approach (see Section 2.2).

CART splits the data into mutually exclusive subgroups, called nodes, within which the objects have similar values for the response variable. The starting point is the root or parent node, which contains all objects of the data set. Then a repeated binary splitting procedure is used to split the data in two groups, called child nodes. The process is repeated by treating each child node as a parent node. Each split is defined by a single explanatory variable and a cut point (for numerical variables) or by relating one or more levels of the (categorical) variable to one of the nodes. For each node all possible splits are evaluated, testing all predictors and their possible threshold values or levels, and finally, the best split is retained. The best split is defined as the variable (and associated splitting value) that minimizes the impurity, i, of the two child nodes. The goodness of a split is then defined as the impurity decrease between the parent node and its children:

$$\Delta i(s, t_P) = i_P(t_P) - p_L i(t_L) - p_R i(t_R) \quad (1)$$

where s is a candidate split; p_L and p_R are the fractions of observations of the parent node t_p that go into the child nodes t_L and t_R, respectively. The best splitter is the one that maximizes $\Delta i(s, t_P)$.

Different criteria to measure the impurity of a node have been proposed for CART [10]. For regression trees, the total sum of squares of the response values about the mean of the

node is the most popular measure of impurity [10,22]:

$$i(t) = \sum_{x_n \in t} (y_n - \bar{y}(t))^2 \quad (2)$$

where $i(t)$ is the impurity of node t; y_n is the response value of observation x_n belonging to node t and $\bar{y}(t)$ the mean of all observations in node t. Absolute deviations about the node medians is another criterion which is used to build (robust) trees [10].

A label or class is assigned to every node of the tree. For regression trees, this is simply the mean within the node.

The splitting procedure is continued until a stopping criterion is reached, i.e. all child nodes are homogeneous, or contain one or a user-defined number of objects. The tree thus obtained is called the maximal tree and describes the training data as good as possible [10]. For this tree, overfitting generally is observed, which will cause poor predictive abilities for new samples [13–15]. However, if one wants to describe the given (training) data set as good as possible, the maximal tree is the best choice. For prediction purposes, the optimal tree is selected from a set of subtrees derived from the maximal tree by means of a so-called pruning procedure. This consists in cutting away the worst terminal branches of the maximal tree. Usually the optimal tree is then selected using either cross-validation methods or based on an external independent test set [10].

Since in this study prediction for new objects was not a goal, the details on pruning and optimal model selection are not described. More information on these steps can be found in references [10,22].

The importance of the explanatory variables to introduce a split in the tree is detected in CART by the variable ranking method as the impurity decrease sum (M_v) caused by a predictor variable v taking into account all nodes of a given tree:

$$M_v = \sum_{t \in T} \Delta i(\tilde{s}_v, t)$$

with $\Delta i(\tilde{s}_v, t)$ the largest reduction in impurity caused by a surrogate split defined by the variable $v(\tilde{s}_v)$ for a node t of the tree T. Thus, the use of each variable in surrogate splits is evaluated for all nodes of the tree. Surrogate splits are alternative splits of a given node, which may be used in the case of missing values for the variable that defines the original split of that node in the tree. The predictor with the largest impurity decrease sum (M_v) is the most important and obtains an importance-value (imp_v) equal to 1. All other variables get a score on the importance scale by comparing their impurity decrease sums relative to that of the most important predictor [10,11,22]:

$$imp_v = \frac{M_v}{\max_v[M_v; \quad v = 1, \ldots, n]}$$

with imp_v the importance of variable v, M_v the impurity decrease sum caused by variable v, and n the number of given variables.

In summary, for each node (t) of a tree (T) the impurity, $i(t)$, is computed as a measure of inhomogeneity. A given split, which divides a parent node into two child nodes, is evaluated based on the impurity decrease (Δi) it causes. A variable (v) is characterized with an impurity decrease sum (M_v), caused by the best (surrogate) splits the variable (v) defines for all splits of a given tree. In the end a variable importance list is obtained after rescaling M_v between 0 and 1, so that the most important variable obtains an importance variable (imp_v) equal to 1.

2.1.2. Relative importance sum

In order to combine the information of a set of univariate regression trees, a new parameter is needed. In other methods that combine several tree-based structures (e.g. boosting CART [23] and random forests [24]), a variable importance parameter is used, that is computed as a weighted sum of the original CART [10] variable importance values for each tree. Here, an analogue parameter, the so-called relative importance sum (RIS) is calculated in order to encode for the general importance of a given predictor variable in a collection of univariate regression trees. This set of trees consists of a separate tree for each of the variables (considered once as response), using the other variables as explanatory variables. Thus, the number of univariate regression trees included in this set equals to the number of variables present in the data set. Using each of the variables as the response of one tree, and the remaining variables as predictor variables to define that tree, the importance (imp) of each variable is computed for all these trees. The relative importance sum is then defined as:

$$\mathrm{RIS}_v = \frac{\sum_{i=1}^{m} imp_{v,i}}{\max_v[\sum_{i=1}^{m} imp_{v,i}; \quad v = 1, \ldots, n]}$$

with RIS_v the relative importance sum of variable v, m the number of univariate regression trees, $imp_{v,i}$ the importance of variable v in the ith tree, and n the number of given variables (here, $n = m$).

2.2. Auto-associative multivariate regression trees

2.2.1. Multivariate regression trees

Whereas univariate regression trees handle only one response, the so-called multivariate regression trees are decision trees that describe several response variables simultaneously. Note that no multivariate decision rules are used for the binary splits in the MRT, but the response of a MRT is multivariate. The construction of an MRT is analogue to a CART, and most of the parameters defined are the same. However, changes are made to the impurity measure and the labelling of the nodes, given the fact that several responses are used [11]. Comparable to CART, a repeated binary splitting procedure is used in MRT to divide the objects into groups (leaves) of analogue response profiles (MRT).

The multivariate impurity measure for MRT is defined as the squared Euclidean distance of objects around the node

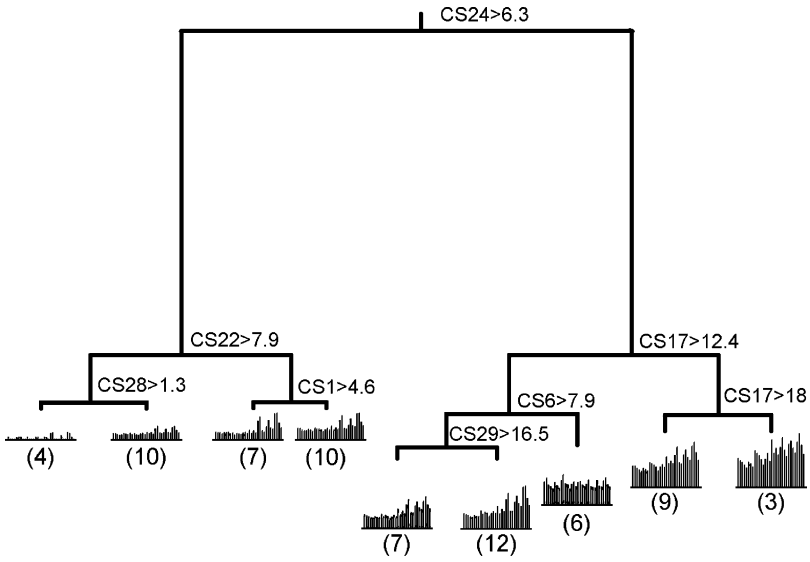

Fig. 1. Multivariate regression tree with nine nodes, describing 32 separate responses. For each node a bar plot represents the distribution of the values of the responses for the objects (molecules) included.

centroid:

$$i(t) = \sum_{x_n \in t} \sum_{j=1}^{p} (y_{n,j} - \bar{y}_j(t))^2$$

where $i(t)$ is the impurity of node t; p the number of response variables described, $y_{n,j}$ the jth response value of observation x_n belonging to node t; $\bar{y}_j(t)$, the mean jth response of all observations in node t [11]. Thus, both in CART and MRT, a decision tree is grown in such a way that the so-called homogeneity and the impurity within each node are maximized and minimized, respectively.

For MRT the mean for each response is the label represented in a bar plot, that shows the distribution of the responses in a given node. The example tree in Fig. 1 defines nine leaves using eight splits. Since one chromatographic system (CS 17) is selected twice to define a split of the tree, a total of seven chromatographic systems is used to grow the tree. Within the nine terminal leaves obtained, the molecules have analogue multivariate response values (here, retention profile on the 32 systems). For each leaf the mean multivariate response profile is represented by a bar plot, in which each bar's height is related to the mean value for each response. The first leaf from the left, for instance, contains four objects with overall very low response values, whereas in the last one 3 molecules with very high responses are grouped.

2.3. Supervised versus unsupervised trees

Recently, Questier et al. [18] proposed a specific kind of multivariate regression trees, called auto-associative multivariate regression trees. These trees differ from the "regular" MRT's, because the predicted (described) variables and the explanatory variables are the same. In general, multivariate regression trees are applied for supervised applications, i.e. to predict a given response using a set of predictor variables, which are different from the multivariate response. Since in AAMRT the same data set is used both as multivariate response and as predictors, the method is unsupervised. Moreover, AAMRTs can only be applied for exploratory data analysis and not for predictions. AAMRT was successfully applied for revealing both clusters and the variables which are most responsible for the cluster structure in a given data set [18]. Since we want to extract the most orthogonal systems for given sets of chromatographic systems based only on the retention data on these systems, AAMRT can be applied.

3. Experimental

The retention data studied were taken from two publications by Van Gyseghem et al. [5,6].

The first data set consists of retention data for 68 structurally diversive drugs (from different pharmacological groups, with different functional groups, pK_a values and hydrophobic properties), on 32 chromatographic systems consisting of eight stationary phases (all silica-based): Zorbax Extend-C18, Zorbax Bonus-RP, Waters XTerra MS C18, Waters XTerra RP18, YMC-Pack C4, Waters SymmetryShield RP18, YMC-Pack Pro C18 and Waters XTerra Phenyl, combined with mobile phases at four different pH values (2.5, 4.8, 7.0 and 9.0). The data set contains mainly basic substances (55), since most drugs have basic properties, but also some neutral (4) and acidic compounds (9) are included. The composition of the different mobile phases, run conditions, and more details on the other chromatographic conditions can be found in [5] for all systems.

The second data set combines a diverse set of 12 columns (including a broad range of stationary phase types): Chromolith Performance, Zorbax Extend-C18, ZirChrom-PS,

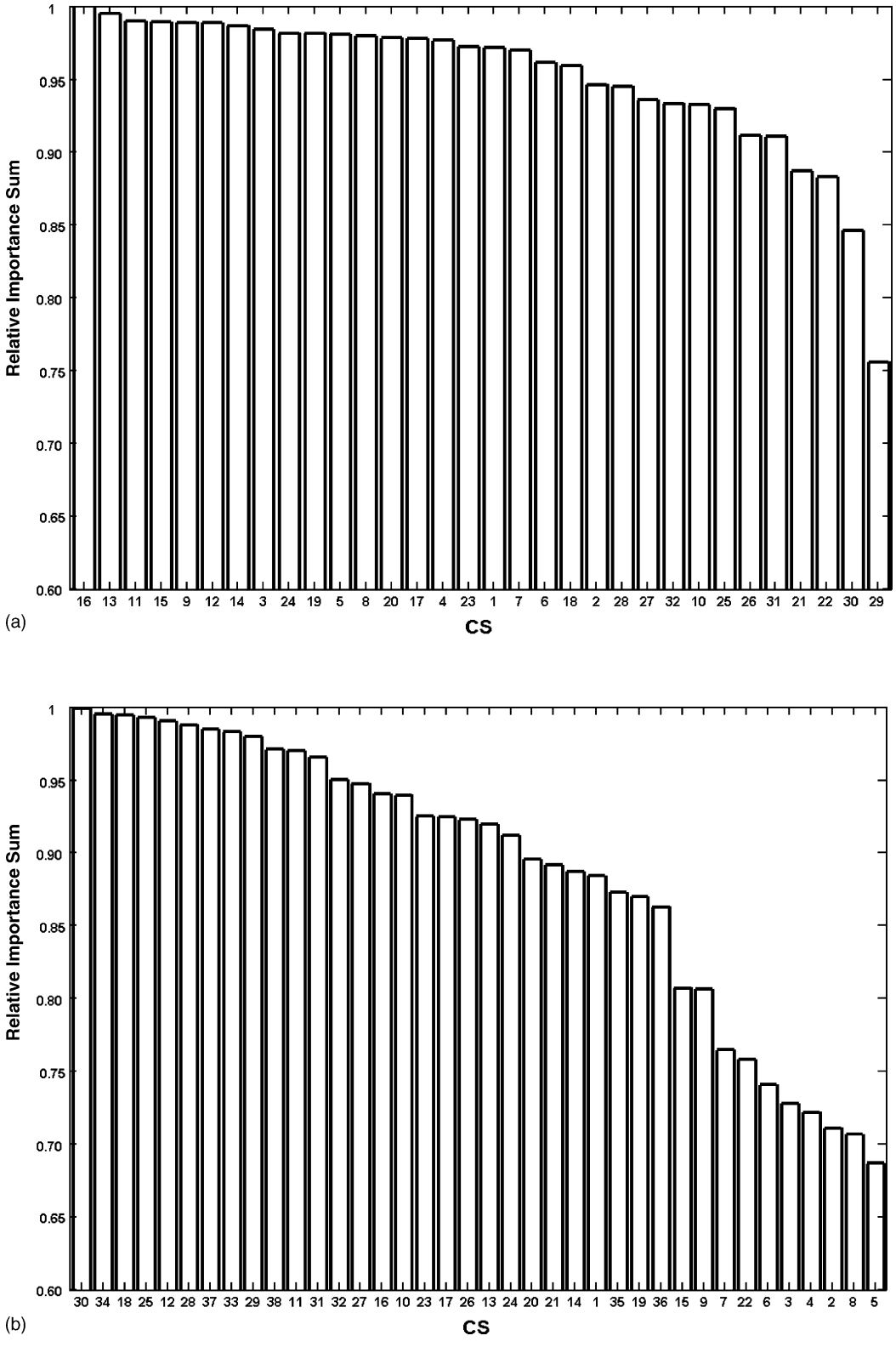

Fig. 2. Relative importance sum (RIS) plot for the (a) 32 chromatographic systems of data set 1, and (b) 38 chromatographic systems of data set 2.

Platinum C18, Platinum EPS C18, Zorbax Eclipse XDB-C8, Betasil Phenyl Hexyl, Suplex pKb-100, ZirChrom-PBD, Aqua, PLRP-S, Luna CN. Combined with mobile phases differing in pH, type of organic modifiers, column temperature and flow rate a total of 38 systems was thus obtained [6]. The retention data used consist of the normalized retention times $\tau = (t_r - t_0)/t_0$ of the 68 substances measured on each of the chromatographic systems using gradient elution. More information on the chromatographic systems can be found in [6].

The tree models were grown using the TreePlus add-on module [25,26] in the S-Plus 2000 environment (Mathsoft, Cambridge, MA, USA) using the following parameters: squared deviations were used as impurity measure, the maximal trees were build with 'one object' as stopping criterion, and all variables were considered for each split of a given tree to define surrogate splits.

4. Results and discussion

In order to define orthogonal systems within the sets of chromatographic systems researched, two approaches were evaluated: the first methodology uses a set of maximal univariate regression trees, each of them build in order to describe the retention on one of the given chromatographic systems (the response), using all remaining retention data as explanatory variables. A total of 32 and 38 univariate regression trees were build for the two data sets, respectively. Based on the computation of the relative importance sum for each of the chromatographic systems a ranking of the chromatographic systems can be obtained.

The second approach is based on the construction of a maximal auto-associative multivariate regression tree, using the normalized retention time τ of the 68 substances on the chromatographic systems (32 and 38, respectively) both as responses and as predictors. Then, the evaluation of the relative importance of the given chromatographic systems in the AAMRT leads to a list of the systems, ranked by their orthogonality (dissimilarity) within the given set of systems. Note that the two data sets were always studied separately.

4.1. Univariate regression trees

4.1.1. Data set 1

Since one univariate regression tree only provides information on the relation between the predictor variables and the response (here, one of the systems), it cannot provide the information needed to select the most orthogonal systems from a given set. However, based on one tree the other columns could be classified by either their similarity or their dissimilarity (orthogonality) regarding to the response column, using the variable importance list of all possible predictors. However, a combination of univariate regression trees (each chromatographic system once as response) can provide more information by summing the individual variable importances obtained for each of the univariate trees. Fig. 2a gives a bar plot of the relative importance sum values obtained for all chromatographic systems, taking into account all 32 univariate regression trees. In this plot the chromatographic systems are ranked according to decreasing RIS value. Thus, all systems are ranked by their decreasing ability to describe the retention of the test substances on the other chromatographic systems, and consequently the most orthogonal systems can be derived as the systems with the lowest RIS values. The 10 most orthogonal systems (lowest RIS values) obtained from Fig. 2a are CS 29, CS 30, CS 22, CS 21, CS 31, CS 26, CS 25, CS 10, CS 32 and CS 27. Analogue systems were defined as orthogonal by Van Gyseghem et al. [5]: the four most orthogonal systems (CS 29, CS 30, CS 22 and CS 21) were also selected in the previous study, together with CS 25, CS 10 and CS 27. Only 3 chromatographic systems (CS 26, CS 31 and CS 32) selected based on their RIS values, were not selected by Van Gyseghem et al. Since none of the methods used can be considered as the only good method to select the most orthogonal systems, it is impossible to conclude which of them is the best. Since overall a high agreement is found between the orthogonal systems selected by these methods, it is concluded that both provide useful information on the orthogonality.

4.1.2. Data set 2

Since 38 different chromatographic systems are included in the second data set, 38 maximal univariate regression trees were build. Based on the importance values obtained for the trees that describe the remaining 37 chromatographic systems, the RIS values were calculated for each of the systems. The resulting bar plot is shown in Fig. 2b. From this plot it can be concluded that CS 5, CS 8, CS 2, CS 4, CS 3, CS 6, CS 22, CS 7, CS 9 and CS 15 are the 10 most orthogonal systems for the second data set. Analogue results are obtained compared to the selection of orthogonal systems from Van Gyseghem

Table 1
Comparison of the most orthogonal systems selected for data set 2: (a) using the univariate approach; (b) based on AAMRT; (c) obtained in [6] (not ordered); (d) obtained in [19]

Orthogonality	(a)	(b)	(c)	(d)
1	CS 5	CS 5	CS 2	CS 5
2	CS 8	CS 2	CS 3	CS 2
3	CS 2	CS 4	CS 4	CS 8
4	CS 4	CS 6	CS 5	CS 3
5	CS 3	CS 8	CS 6	CS 6
6	CS 6	CS 3	CS 7	CS 4
7	CS 22	CS 15	CS 8	CS 7
8	CS 7	CS 7	CS 9	CS 22
9	CS 9	CS 22	(CS 15)	CS 9
10	CS 15	CS 9	CS 20	CS 15
11	CS 36	CS 1	CS 22	CS 20
12	CS 19	CS 19		CS 36
13	CS 35	CS 14		CS 14
14	CS 1	CS 20		CS 1
15	CS 14	CS 13		CS 19

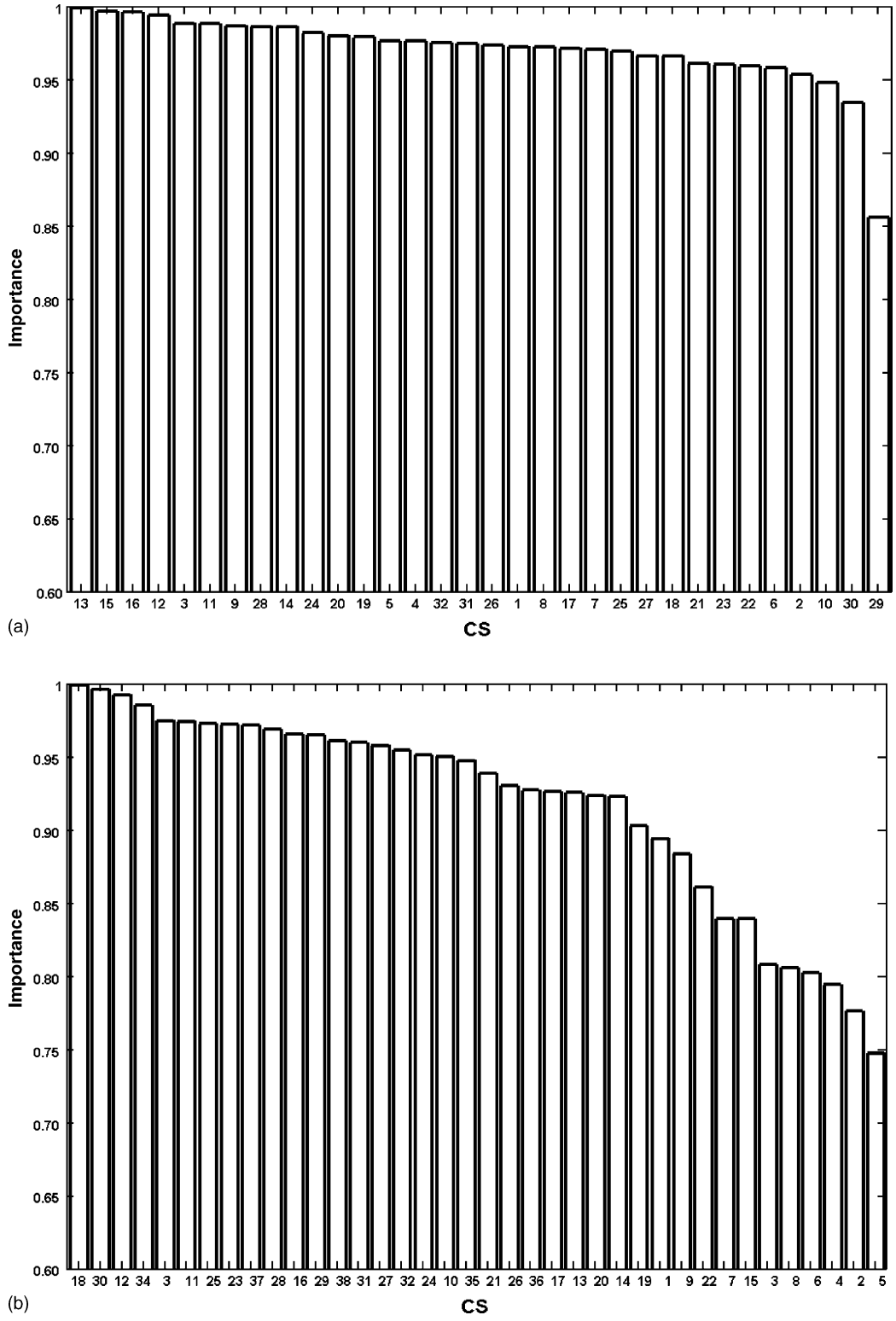

Fig. 3. Importance plot for (a) all predictors (32 chromatographic systems) of data set 1, (b) all predictors (38 chromatographic systems) of data set 2.

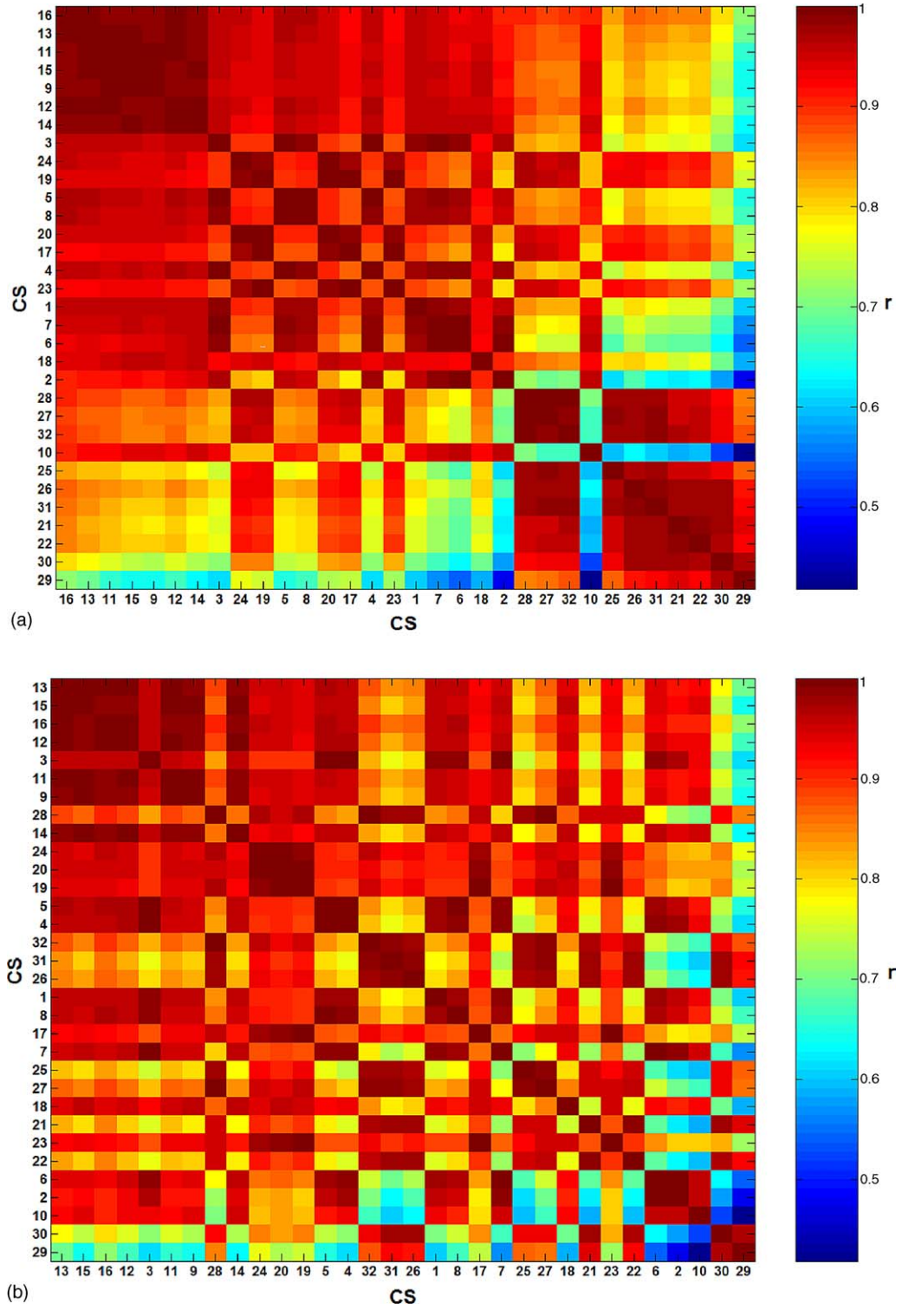

Fig. 4. Color map for the matrix of Pearson correlation coefficients r between the normalized retention times τ of the 32 systems of data set 1. The systems were ranked according to (a) their RIS value calculated for the 32 univariate regression trees, (b) their importance in the AAMRT.

et al. [6]. All systems selected here, were also selected as orthogonal based on the evaluation of the correlation coefficients using dendrograms and color maps. In the paper by Forlay-Frick et al. [19], where selection is based on the generalized pairwise correlation method with the McNemar's test the same 10 systems were selected (Table 1). Moreover, from their 15 most orthogonal ones, only CS 20 is not defined as orthogonal based on the RIS value.

When comparing Fig. 2a and b one observes that the relative importance sum differences are more important in Fig. 2b. The RIS-values of the orthogonal systems in Fig. 2b also are considerably smaller than those in Fig. 2a. This can be explained from a practical point of view. In data set 1 (Fig. 2a) we are dealing with a set of relatively similar reversed-phase stationary phases evaluated at different pH values. In data set 2 (Fig. 2b) stationary phases with very different properties are included. Therefore, one can expect that the selectivity differences between the most diverse systems of data set 2 will be larger than between those of data set 1. This is reflected in the RIS plots.

4.2. Auto-associative multivariate regression trees

4.2.1. Data set 1

A maximal auto-associative multivariate regression tree was build to describe the retention on the 32 chromatographic systems. In general, multivariate regression trees tend to split the data into groups with analogue multivariate response profiles. As discussed by Questier et al. [18] AAMRTs can be useful to uncover the most important variables in complex high-dimensional data sets. One could suggest that a chromatographic system selected to define a split in the tree can be considered as a "general" chromatographic system, since it can be used to describe the retention profile of a molecule, which means the chromatographic retention on several systems. Moreover, the most general systems (i.e. these with similar selectivities) will perform better to define both the primary splits, and the surrogate splits of the tree. As a consequence, a chromatographic system that has a completely different selectivity (and thus is orthogonal) will be less important to define the tree, since it does not contain information related to the other systems. Thus, the chromatographic systems that contribute the least to the AAMRT can be defined as most orthogonal to the rest. Fig. 3a shows a bar plot in which the chromatographic systems are ranked according to their importance in the global AAMRT. The system with the largest importance is situated on the left of the plot and its bar height equals 1. The bars of the other systems (with lower importance values) are sorted by decreasing importance. From this ranking it can be concluded that the 10 most orthogonal systems are the systems CS 29, CS 30, CS 10, CS 2, CS 6, CS 22, CS 23, CS 21, CS 18 and CS 27, since they show the lowest importances in the tree. Four of the 10 systems selected as orthogonal based on the AAMRT approach, were not selected using the univariate approach. However, three of them were also selected in the work of Van Gyseghem et al. [5]: only CS 23 was not. The differences between the selections based on different approaches might be explained by

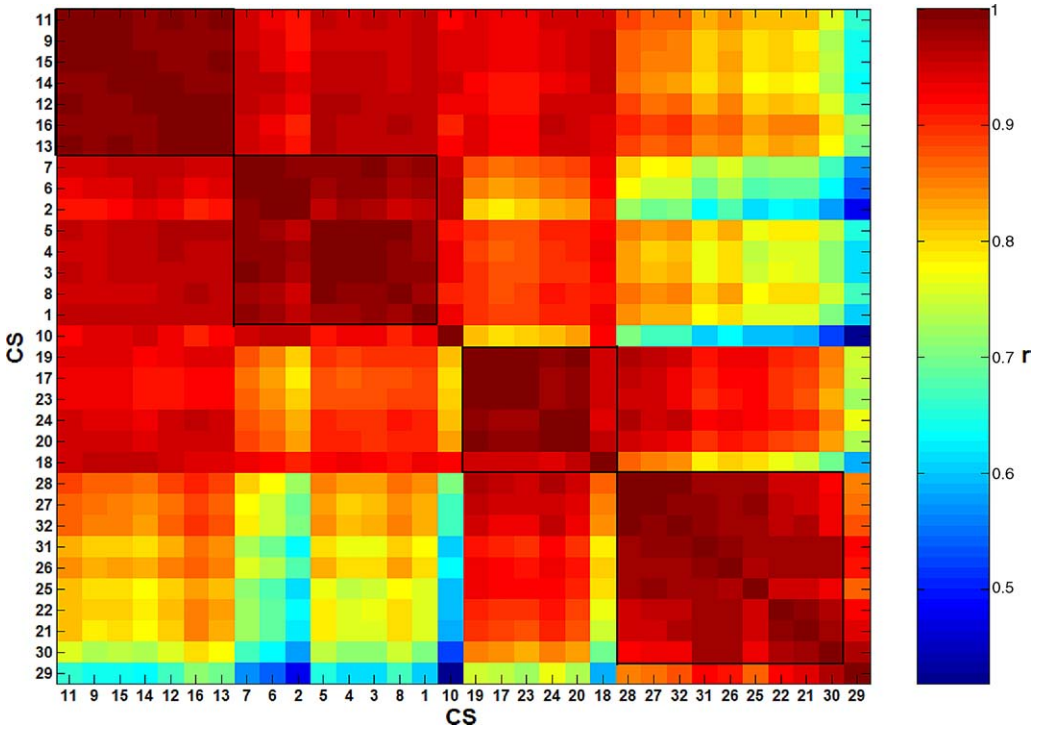

Fig. 5. Color map of correlation coefficients for the 32 systems ranked according to increasing dissimilarities in the weighted-average-linkage dendrogram (redrawn from data set of [5]).

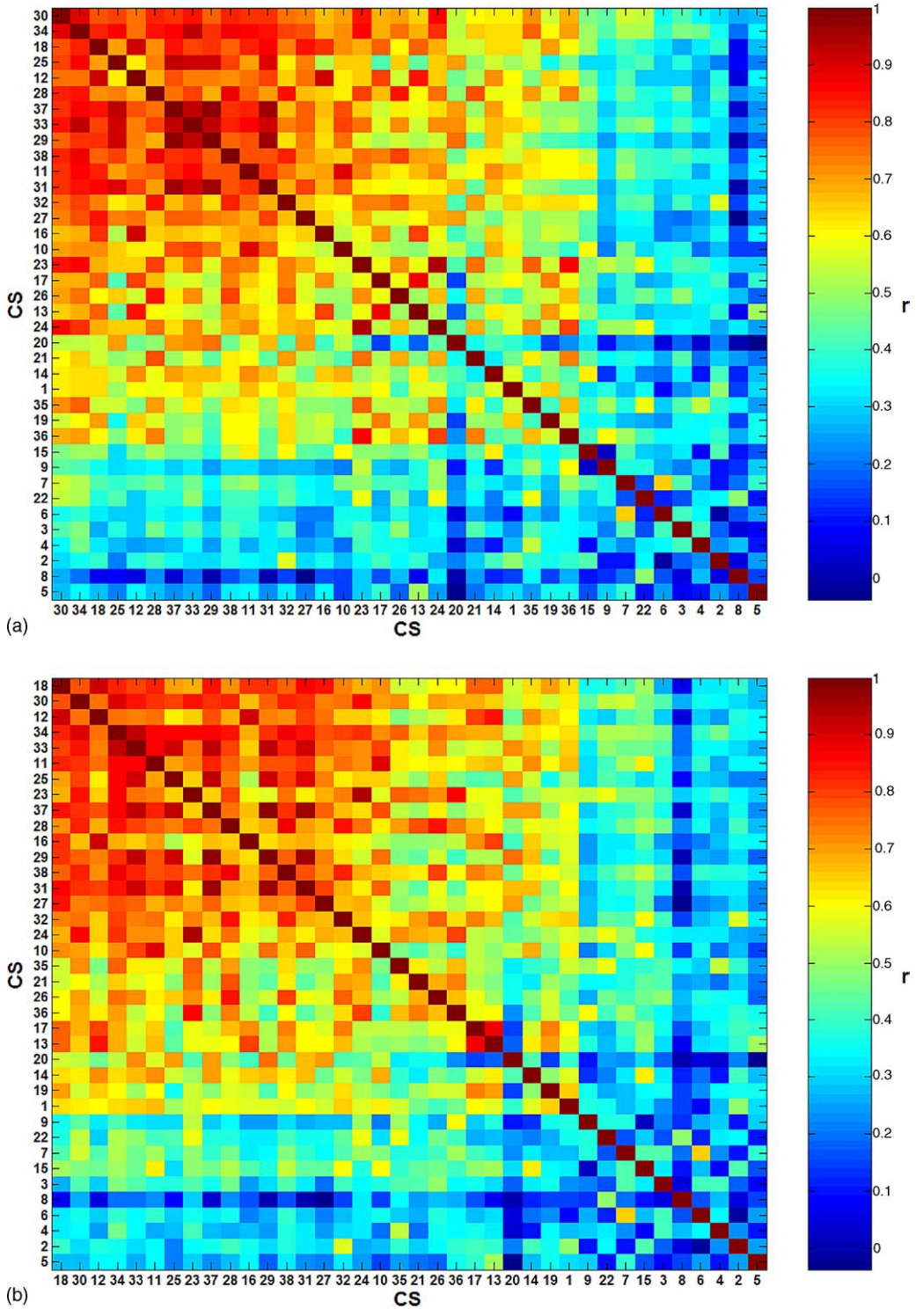

Fig. 6. Color map for the matrix of Pearson correlation coefficients r between the normalized retention times τ of the 38 systems of data set 2. The systems were sorted according to (a) their RIS value calculated for the 38 univariate regression trees, (b) their importance in the AAMRT.

the higher discussed relative similarity between the different systems.

4.2.2. Data set 2

The AAMRT results in the variable importance list are shown in Fig. 3b. From this plot it can be concluded that the same chromatographic systems are considered to be orthogonal from the univariate approach based on the RIS parameter. The following systems were thus most orthogonal for this data set: CS 5, CS 2, CS 4, CS 6, CS 8, CS 3, CS 15, CS 7, CS 22 and CS 9. For AAMRT analogue results are obtained compared to the univariate approach. In spite of some shifts in the ranking, the same 10 most orthogonal systems were selected for the second data set. In the list of the 15 most orthogonal systems (Table 1), CS 35 and CS 36 are here replaced by CS 20 and CS 13. Compared to the study by Forlay-Frick et al. [19] now only system 36 is not selected and is replaced by system 13 in the list of the 15 most orthogonal systems, while CS 20 however, is included. Table 1 indicates that in general the different methods lead to an analogue selection of the most orthogonal chromatographic systems from the 38 systems. However, the exact sequence of orthogonality is not defined uniformly, since the basis on which orthogonality is defined differs for all methods. The generally accepted definition of orthogonal systems states that a considerable selectivity difference between such systems is observed [5,27], but there is no preferred mathematical methodology yet for the selection of the most orthogonal systems. As a consequence, it may be advisable to consider several methods simultaneously in order to make such a selection, instead of using only one (preferred) method. Taking this into account, the preferable set of orthogonal systems is formed by CS 2, CS 3, CS 4, CS 5, CS 6, CS 7, CS 8, CS 9, CS 15 and CS 22, since these systems are selected by each method considered.

When comparing the importance plots for both data sets (Fig. 3a and b) similar conclusions can be drawn as when the RIS plots (Fig. 2a and b) were.

4.3. Color maps

The use of color maps, as an additional visualization technique to evaluate, for instance, groups of similar systems, was examined. Color maps were build as proposed by Van Gyseghem et al. [5,6], representing the Pearson's correlation coefficients (r) calculated between the normalized retention time τ of the substances on each pair of systems. Here, the ranking of the chromatographic systems however, is based on the RIS values and importance values of the chromatographic systems obtained by the univariate regression trees and auto-associative multivariate regression trees, respectively.

Fig. 4 shows the color maps obtained for data set 1. As can be observed, based on these color maps it is harder to distinguish between groups of similar systems than those based on a ranking according to the weighted average linkage dendrograms, which were applied in [5] (Fig. 5), since the dendrogram tends to group similar objects. Analogue conclusions can be drawn for Fig. 6, which shows the color maps for data set 2. Comparing the Figs. 4 and 6 again demonstrate the larger selectivity differences (reflected in larger r-value differences) in the latter figure. Figs. 4 and 6 indicate that the ranking of the systems in the correlation coefficients matrix according to the results of the regression trees is less appropriate to consider groups of systems in the data set. However, it can be seen that the most orthogonal systems selected, show a low correlation to most other systems (blue colors in the right and lower part of the color maps), which confirms the discussed selection of orthogonal systems.

5. Conclusions

Two new ranking methods were derived for the selection of the most orthogonal chromatographic systems from a given set of systems, which are based on univariate regression trees and auto-associative multivariate regression trees, respectively. The information regarding the most orthogonal systems can be extracted from the univariate regression trees by describing the retention on a chromatographic system using the retention data on the other systems as predictors. The proposed relative importance sum parameter is used to quantify the information gained. AAMRT creates a simple tree, which divides the molecules into groups within which the objects have similar retention profiles on the different chromatographic systems. The most orthogonal chromatographic systems can be easily derived, using the importance plot of the different chromatographic systems for their use as predictors in the AAMRT. The most orthogonal systems have the lowest importance, since they have the least in common with the other chromatographic systems and thus are the most different from them. Compared to previous studies, similar selections were made: e.g. for the second data set exactly the same 10 most orthogonal systems were found with both the univariate approach and AAMRT, which were also equal to an earlier selection based on statistical tests.

References

[1] P. Jandera, in: K. Valko (Ed.), Handbook of Analytical Separations, vol. 1, Elsevier, Amsterdam, 2000, p. 1, Chapter 1.
[2] R.J.M. Vervoort, A.J.J. Debets, H.A. Claessens, C.A. Cramers, G.J. de Jong, J. Chromatogr. A 897 (2000) 1.
[3] H.A. Claessens, TrAC Trends Anal. Chem. 20 (2001) 563.
[4] A. Detroyer, V. Schoonjans, F. Questier, Y. Vander Heyden, A.P. Borosy, Q. Guo, D.L. Massart, J. Chromatogr. A 897 (2000) 23.
[5] E. Van Gyseghem, M. Jimidar, R. Sneyers, D. Redlich, E. Verhoeven, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 1074 (2005) 117.
[6] E. Van Gyseghem, I. Crosiers, S. Gourvénec, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 1026 (2004) 117.
[7] U.D. Neue, B.A. Alden, T.H. Walter, J. Chromatogr. A 849 (1999) 101.
[8] S.M. Fields, C.Q. Ye, D.D. Zhang, B.R. Branch, X.J. Zhang, N. Okafo, J. Chromatogr. A 913 (2001) 197.

[9] U.D. Neue, E.S. Grumbach, J.R. Mazzeo, K. Tran, D.M. Wagrowski-Diehl, in: I.D. Wilson, A. Macclesfield (Eds.), Handbook of Analytical Separations, vol. 4, Elsevier, Amsterdam, 2003.
[10] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth, Monterey, CA, USA, 1984.
[11] G. De'ath, Ecology 83 (2002) 1105.
[12] R. Put, C. Perrin, F. Questier, D. Coomans, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 988 (2003) 261.
[13] N. Lavrač, Artif. Intell. Med. 16 (1999) 3.
[14] R.J. Marshall, J. Clin. Epidemiol. 54 (2001) 603.
[15] G. De'Ath, K.E. Fabricius, Ecology 81 (2000) 3178.
[16] M.R. Segal, J. Am. Stat. Assoc. 87 (1992) 407.
[17] D.R. Larsen, P.L. Speckman, Biometrics 60 (2004) 543.
[18] F. Questier, R. Put, D. Coomans, B. Walczak, Y. Vander Heyden, Chemom. Intell. Lab. Syst. 76 (2005) 45.
[19] P. Forlay-Frick, E. Van Gyseghem, K. Heberger, Y. Vander Heyden, Anal. Chim. Acta 539 (2005) 1.
[20] E. Van Gyseghem, S. Van Hemelryck, M. Daszykowski, F. Questier, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 988 (2003) 77.
[21] K. Héberger, R. Rajkó, J. Chemom. 16 (2002) 436.
[22] R. Put, C. Perrin, F. Questier, D. Coomans, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 988 (2003) 261.
[23] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, New York, USA, 2001.
[24] L. Breiman, Random forests, Technical Report, University of California, Berkeley, 2001.
[25] G. De'ath, Ecological Archives E083-017-S1, 2002, http://www.esapubs.org/archive/ecol/E083/017/.
[26] G. De'Ath, Ph.D. thesis, James Cook University, Townsville, Australia, 1999.
[27] G. Xue, A.D. Bendick, R. Chen, S.S. Sekulic, J. Chromatogr. A 1050 (2004) 159.