ELSEVIER

Analytica Chimica Acta 348 (1997) 357–374

ANALYTICA CHIMICA ACTA

Discrimination of the variety and region of origin of extra virgin olive oils using ^{13}C NMR and multivariate calibration with variable reduction

Adrian D. Shaw[a,*], Angela di Camillo[b], Giovanna Vlahov[b], Alun Jones[a], Giorgio Bianchi[b], Jem Rowland[c], Douglas B. Kell[a]

[a]*Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion SY23 3DA, UK*
[b]*Istituto Sperimentale per la Elaiotecnica, Contrada "Fonte Umano" n 37, 65013 Città S. Angelo, Pescara, Italy*
[c]*Department of Computer Science, University of Wales, Aberystwyth, Ceredigion SY23 3DB, UK*

Received 1 June 1996; received in revised form 2 December 1996; accepted 31 December 1996

Abstract

There is strong evidence that consumption of olive oil, especially extra virgin olive oil, reduces the risk of circulatory system diseases. Such oil is generally more expensive than other edible oils, Italian — and in particular Tuscan — oils being particularly favoured by connoisseurs, and commanding an even higher price. There is therefore a great temptation to adulterate olive oil with a cheaper oil, or falsify its origin or grade. An easy and reliable method to identify different types of olive oil is required. Our work has focused on discriminating extra virgin olive oils by their region and variety. We have applied Principal Components Analysis (PCA), Principal Components Regression (PCR) and Partial Least Squares (PLS) to discriminate olive oils on the basis of their ^{13}C NMR spectra. *Variable Selection* was used in order to reduce the number of variables in the data. Two main methods of variable selection have been used; these are the Fisher Ratio, and the ratio of *Inner Variance* to *Outer Variance* or *Characteristicity* [W. Eshuis, P.G. Kistemaker and H.L.C. Meuzelaar, in C.E.R. Jones and C.A. Cramers (Eds.), Analytical Pyrolysis, Elsevier, Amsterdam, 1977, pp. 151–156.]. Both these methods proved successful in improving the PCA clustering, and the prediction results of PCR and PLS, although the optimal number of variables varied between datasets. PCR2 and PLS2 models, in which a single model is used to predict each variety or each region simultaneously, achieved a successful prediction rate of some 70%. However, multiple PLS1 models routinely achieved successful predictions of over 90% and in many cases 100% of the data in test sets. Indeed the variety of all but 1 of 66 samples was correctly predicted. It is clear that multiple, specialised models perform much better than "global" ones, and that the inclusion of certain variables can be highly detrimental to the multivariate calibration process.

Keywords: Olive oil; Adulteration; Chemometrics; PLS; Variable reduction

1. Introduction

The value of olive oil produced annually is around $2.5 billion [1], other olive products amounting to

*Corresponding author. Tel.: +44 1970 622334; fax: +44 1970 622354; e-mail: ais@aber.ac.uk.

PII S0003-2670(97)00037-8

around 300 million dollars. 9.4 million tonnes of olive fruit is produced per annum, from 805 million olive trees worldwide, occupying some 24 million acres of land. 98% of these trees are in the Mediterranean area. Of the 60 million tonnes of seed oil consumed worldwide every year, 2 million are olive oil [2].

Almost 25% of the farming income in the Mediterranean basin as a whole comes from olive products, Spain and Italy being far and away the largest producers, with Greece (with around half the production of the main two) coming third. In 1987, Italy contributed about 65% of world olive oil [3].

Virgin olive oil is the oil extracted by purely mechanical means from sound, ripe fruits of the olive tree (*Olea europaea L.*). Extra virgin olive oil is absolutely perfect in flavour and odour, and has a maximum free fatty acid content in terms of oleic acid of 1 g/100 g [1,4].

Compared to other edible oils, olive oil contains a low percentage of saturated fatty acids (that is, fatty acids with no double bonds in the carbon chain) at around 16% (mainly 16 : 0). It contains a high percentage of monounsaturated fatty acids, around 70% (mainly 18 : 1) and around 15% polyunsaturated fatty acids (18 : 2) - where 18 : 2 indicates a chain of 18 carbon atoms, with two unsaturated (C=C) bonds, etc.

Olive oil has a fine aroma and a pleasant taste, which is generally agreed to be at its best in extra virgin olive oils, and is considered to have many nutritional and health benefits [1].

There are many varied claims and suggested reasons as to the health benefits. There is very strong evidence that olive oil consumption reduces the risk of death due to circulatory system diseases [1,5]. Visioli [6] and Galli [7] suggest that this is due at least partially to the natural antioxidants (including the bitter-tasting glycosidic compound Oleuropein) and micronutrients preventing low density lipoprotein from oxidation and so retarding the formation of atherosclerotic lesion.

Martin-Moreno et al. [8] also note that olive oils contain a "generous amount of antioxidants" and speculate that "diets high in monounsaturated fats presumably yield tissue structures that are less susceptible to antioxidative damage than would be the case in high polyunsaturated diets". They identify an inverse correlation between breast cancer and olive oil intake, as do Trichopoulou et al. [9]. The latter also claim that margarine consumption increases this risk. Trichopoulou et al. [10] suggest that olive oil consumption is one of the factors in the traditional Greek diet that aids the longevity of those elderly people in a study group who followed that diet.

As a consequence of these benefits, olive oil commands a much higher price than most other edible oils. This in turn means that there is a great temptation to adulterate the oil with a cheaper oil, such as olive pomace oil, corn oil, sunflower oil, or even lard or castor oil [11–13]. In addition, many oils labelled as "extra virgin" have been processed in order to reduce the acidity level and so gain this classification.

Italian olive oil, and in particular Tuscan olive oil, is traditionally the most favoured, and therefore attracts a premium. It is not surprising, then, that there is more oil labelled as Tuscan than could possibly be produced there. Firestone et al. [12] reported on a US survey in which 4 out of 5 virgin olive oils were correctly labelled, but only 3 out of 20 olive oils. In 1988, they followed up their 1985 report [13], noting some improvement. This time, although only 17 out of 31 virgin olive oils were correctly labelled, so were 15 out of 26 olive oils; still over 40% incorrectly labelled.

The necessity to be able to detect adulterations in oils in general was highlighted in May 1981, when 20 000 people became ill and 350 died in Spain after consuming oils containing "refined" aniline denatured rape seed oil [14].

A number of researchers have applied themselves to this area.

Grob et al. [15] report that extra virgin oils (known as cold pressed or non-refined when referring to non-olive oils) can be distinguished by the presence of a substantial quantity of volatile components (i.e. they have not been deodorised). If none of these volatiles are present, the oil has been treated. "Pure" oils, being a blend, are more difficult to distinguish. Grob et al. [16] were able to detect adulteration of olive oils down to 10% (even lower for most oils) using Liquid Chromatography–Gas Chromatography–Flame Ionisation Detection (LC–GC–FID) by direct analysis of these minor components. They do note however that strong raffination made adulteration difficult to detect.

A variety of workers have shown that it is largely possible to discriminate seed oils on the basis of their fatty acid content as judged by GC [17–20], whilst we [4,21] were successful in detecting adulteration of

extra virgin olive oil using pyrolysis mass spectrometry and Artificial Neural Networks (ANNs).

Sato [22] showed that near infrared spectroscopy can be used with PCA to discriminate many vegetable oils from each other, including olive oil. Schwaiger and Vojir [23] had similar success with GC analysis and PCA, the first two principal components separating olive oil well from the other oils.

Zamora et al. [11] were able to distinguish between different grades of oil using ^{13}C NMR, whilst Lai et al. [24,25] have also successfully applied the technique of FTIR for assessing the authenticity of vegetable oils.

One advantage of the spectroscopic methods over many other methods is that it is possible to look at the whole spectrum and search for patterns emerging from the data [26]. No information is thrown away; rather any element in the spectrum which appears to indicate the presence of adulterants will be given a larger weighting in the model. Work done by others also supports this as the most suitable approach, since they fully utilise even subtle variations [27]. Indeed, this type of approach has been previously used for crude petrochemical oil, as described by Kvalheim et al. [28] and Brekke et al. [29], with very promising results.

In the present work we show for the first time that the ^{13}C NMR spectra of oils from different regions and of different varieties are sufficiently different to permit their discrimination using advanced chemometrics methods.

2. Materials and methods

2.1. NMR

Olive fruits of varieties Coratina (7), Dritta (1), Grossa di Cassano (3), I-77 (8), Moraiolo (8) and Picholine (4) were sampled in different regions of Italy: Abruzzo (6), Calabria (3), Lazio (1), Lombardia (1), Marche (1), Molise (2), Puglia (7), Toscana (6) and Umbria (3), as well as a sample from four locations in Israel of unknown variety. All these samples were divided into two, thus producing twice the number of objects; the exception being the Dritta sample, which was divided into 12 so that the reproducibility of the method could be ensured. The fruits were processed for oil extraction by a micro-oil-mill within two days of harvesting [4]. The oils were stored at a temperature of −18°C until ^{13}C NMR spectra were registered.

2.2. High resolution ^{13}C NMR

^{13}C NMR spectra were obtained using a Bruker AC 300 spectrometer at the Istituto Sperimentale per la Elaiotecnica.

The spectra of oil samples were run in $CDCl_3$ (200 mg/0.5 ml). Free induction decays (FIDs) were acquired at 25°C using a spectral width of 13 000 Hz, with 131 k acquisition points zero filled to 256 k points.

A 45° excitation pulse and a 20 s relaxation delay were employed to collect 256 scans.

FIDs were processed before Fourier transformation by a Gaussian filter of 0.1 Hz with Lorentzian narrowing and 0.15 Gaussian broadening.

The T1 relaxation times were measured by the inversion-recovery T1 pulse sequence. T1 experiments were run over a restricted spectral width of 2700 Hz, including only the methyl and methylene envelope, with 32 k data points, zero filled to 64 k points to improve digital resolution up to 0.08 Hz/point.

T1 experimental spectra were resolution enhanced by a Gaussian transformation.

The areas under the resonance lines were measured by means of integration using the spectrometer computer.

The following precautions necessary to ensure reliable integration measurements were adopted.

In order to avoid signal saturation, smaller flip angles (45°) and waiting times of 20 s between pulses were used which fulfil the requirement of being 5*T1 necessary to recover all magnetisation between pulses. The effect of differential nuclear Overhauser enhancements operating on carbon-13 nuclei was quenched by using an inverse-gated proton decoupled arrangement.

The most serious difficulty to overcome in signal integration is the definition of the baseline level of the spectrum. This problem was alleviated by carefully limiting the frequency range of integration to the peaks of interest. These limits were kept constant throughout the spectrum integration of different oil samples.

2.3. ^{13}C NMR spectral data and assignments

Carbonyl region. The oils exhibited the resonances of carbonyl carbons C1 of acyl chains in the ^{13}C NMR spectral region in the chemical shift range of 173.2–172.5 ppm (see Table 1).

The signals were readily assigned by comparison with the chemical shifts of standard glycerides and further confirmed by numerous data available in the literature [30].

The C1 resonances were grouped into two well-resolved sets of resonances. The low field resonances comprised the carbonyl signals of saturated (δ=173.11 ppm), oleyl (δ=173.08 ppm) and linoleyl (δ=173.08 ppm) chains esterified at the 1,3-positions of the glycerol backbone.

The high field group of resonances showed the carbonyls of oleyl (δ=172.69 ppm) and linoleyl (δ=172.68 ppm) chains at 2-position of glycerol.

Saturated acid components such as the palmitic acid ($C_{16:0}$) and stearic acid ($C_{18:0}$) chains are not detected at 2-position of glycerol, because in vegetable oils saturated chains are found esterified at that position in very low amount (less than 2%). A 300 MHz NMR fails to detect signals for components present in a mixture at such low levels.

Other results showed that ^{13}C NMR spectra of fairly simple triglyceride mixtures made up of the common C-18 acyl chains, like those of olive oil, assures the separation of the different chains present and the discrimination of their attachment position on glycerol.

Olefinic Region. The signals of double bond carbons of oleyl and linoleyl chains at the 1,3- and 2-positions (130.3–127.7 ppm) of the glycerol backbone were detected and found to be in agreement with literature data and accordingly assigned [31,32].

The most remarkable spectroscopic feature was that the unsaturated carbons of oleyl and linoleyl chains resonate as doublets whose splittings were found whether they were at the 1,3- or 2-positions of glycerol.

Aliphatic Region. The complete assignments of signals in the aliphatic region (35–6 ppm), with identification of acyl chains and their position on the glycerol backbone, were achieved by T1 longitudinal relaxation times.

The T1 values proved highly useful in determining carbon chemical shifts of the methylene envelope. In fact the polar head group slightly perturbs the electronic environment of methylene sp^3 carbons, thus making CH_2 resonances spread over a very narrow range of frequencies.

T1 ^{13}C relaxation times were found to increase regularly from the polar head group along the aliphatic fatty acid chains in $CDCl_3$ [33]. This was interpreted according to the increasing mobility of the chain with distance from the polar group because carboxyls are kept immobilised by a self-association mechanism. The T1 value patterns of saturated, oleyl and linoleyl chains fitted this rule perfectly.

2.4. Chemometrics

The data obtained are arranged in the form of a Microsoft Excel spreadsheet. There was a total of 80 objects, including 12 from the same sample for the purpose of verifying reproducibility. The remaining 68 objects were duplicates, from 34 samples. The data were normalised, so that the total of the integrated

Table 1
Assignments and selection order of the carbon signals used

Region	Carbon signal	Assignment	Chain	Chemical shift (ppm)		Variety selection order			Region selection order		
				α	β	Fisher	Wtd w	Unw w	Fisher	Wtd w	Unw w
Carbonyl	CS_1	C1	S	173.113		14	14	8	18	19	22
	CS_1.1	C1	E	173.101		38	38	39	41	41	41
	CS_2	C1	O	173.084		2	2	2	6	7	4
	CS_3	C1	L	173.075		26	27	27	34	34	33
	CS_4	C1	O		172.688	1	1	1	12	16	15
	CS_5	C1	L		172.679	5	6	15	16	14	17

(*continued overleaf*)

Table 1
(continued)

Region	Carbon signal	Assignment	Chain	Chemical shift (ppm)		Variety selection order			Region selection order		
				α	β	Fisher	Wtd w	Unw w	Fisher	Wtd w	Unw w
Olefinic	CS_6	C13	L	130.097	130.105	17	11	12	27	17	12
	CS_7	C10	O		129.945	34	34	37	26	28	30
	CS_8	C10	O	129.930		29	29	29	23	26	27
	CS_9	C9	L	129.900		28	26	30	24	24	24
	CS_10	C9	L		129.874	12	16	19	20	18	18
	CS_11	C9	O	129.640		20	19	16	1	2	2
	CS_12	C9	O		129.614	27	28	26	29	30	26
	CS_13	C10	L		128.053	10	9	11	25	21	16
	CS_14	C10	L	128.035		13	13	18	14	13	10
	CS_15	C12	L	127.874		18	17	23	13	12	13
	CS_16	C12	L		127.862	16	20	20	19	20	21
Glycerol	CS_17	GL			68.8852	36	36	33	40	40	36
	CS_18	GL		62.0478		31	30	28	28	27	31
Aliphatic	CS_19	C2	S	33.9972		39	40	38	36	37	37
			O	33.9758	34.1408						
			L	33.9758	34.1408						
	CS_20	C16	S	31.9283		19	23	25	32	32	32
			O	31.9074	31.9074						
			L	31.5167	31.5167						
	CS_21	C12	O	29.7586	29.7586	4	5	7	4	5	5
	CS_22	Not known				3	4	4	3	3	3
	CS_23	C7	S	29.6637		7	7	6	5	4	6
	CS_24	C7	L		29.6203	22	21	9	11	15	19
	CS_26	C14	O	29.5289	29.5289	15	15	14	10	9	11
	CS_27	C6	S	29.4748		23	22	17	17	23	23
	CS_28	C15	S	29.3688		9	8	5	7	6	7
	CS_29	C15	L	29.3419	29.3419	11	12	13	22	22	20
	CS_30	C15–C13	O	29.3156	29.3156	6	3	3	2	1	1
	CS_31	C5	S	29.2686		24	24	21	33	33	29
	CS_32	C5	O+L		29.1869	30	31	31	21	25	25
	CS_33	C5	O+L	29.1656		25	25	22	8	11	14
	CS_34	Not known				40	39	40	39	39	40
	CS_35	C4	O+L	29.0717		32	33	32	31	31	34
	CS_36	C4	O+L		29.0324	41	41	41	38	36	39
	CS_37	C8	O	27.1477	27.1477	8	10	10	9	8	9
		C11	O	27.2009	27.2009						
		C8	L	27.1657	27.1657						
		C14	L	27.1815	27.1815						
	CS_38	C11	L	25.6084	25.6084	21	18	24	15	10	8
	CS_39	C3	S	24.8466		35	32	34	30	29	28
			O	24.8239	24.8605						
			L	24.8239	24.8605						
	CS_40	C17	S	22.6841		33	35	35	35	35	35
			O	22.6743	22.6743						
			L	22.5656	22.5656						
	CS_41	C18	S	14.0824		37	37	36	37	38	38
			O	14.0769	14.0769						
			L	14.0370	14.0370						

S=Saturated, O=Oleic, L=Linoleic, E=Eicosenoic.

resonances of the 41 variables for each object added up to 100.

The results were obtained by using Excel 5 macros written specifically for the purpose, in conjunction with PCA, PLS, PCR and MLR programs written in-house. For PLS and PCR, cross-validation was performed using test set - training set validation. All duplicates were kept in the same set, except for the sample used for verifying reproducibility (which was split between the two).

3. Theory: Variable selection

When large amounts of data are passed into a statistics package, or processed by neural networks in order to create some kind of calibration model, some variables in the data will be found to be of greater value than others. Indeed, in the extreme case, one variable might be sufficient to discriminate all the different types of object being examined, whilst other variables contain only noise.

It may be argued that there is never any point in reducing the number of variables before creating a multivariate calibration model, as this is unnecessarily reducing the amount of data that can be used to create the model. It is argued that if any of the variables are of little value (or exhibit collinearity with others), this will be reflected in the weighting given to them in the final model. This view has in recent years been discredited by a number of researchers in both spectrometry [34–51] and QSAR [52–59]. Thus, Brown [42] showed how variable reduction from near infrared spectral data greatly improved the PLS prediction of the amount of sucrose, fructose and glucose in aqueous solution, whilst Sreerama and Woody [47] showed that variable reduction greatly improves the results of PCA for the prediction of protein secondary structure fractions from circular dichroism spectra. Other workers have shown that if a model may be described adequately by different numbers of variables, then that described by the fewest will be able better to generalise [60,61]. Indeed, the danger of obtaining chance correlations is well known to increase with the number of independent variables [62].

Chatfield [63] warns of the dangers of selecting variables in such a way as to enable some sort of model to be made from pure noise. We do not believe this an issue with the present data (and note that we are here applying non-parametric methods). This is because we are creating only one model for each best n variables, not taking the best model from many (our variable selection is not influenced by the model created). Similarly, we have analysed various datasets, and have found the order of selection for variety identification to be similar each time; the same is true for the order of selection for region identification.

Three methods of variable selection have been used. The first is shown in Eq. (1) (for unweighted w):

$$w = \{\text{Average}[\text{StDev}(\text{Variety 1}), \text{StDev}(\text{Variety 2}) \ldots \text{StDev}(\text{Variety } n)]\}/ \text{StDev}(\text{All samples}). \quad (1)$$

In most datasets, it will be found that the number of samples in each variety (or whatever) is not the same. In this case, it may be desirable to weight the value of w in favour of those varieties with more samples (for weighted w, see Eq. (2)):

$$w = [(\text{StDev}(\text{Variety 1}) \times g_1) + (\text{StDev}(\text{Variety 2}) \times g_2) + \ldots + (\text{StDev}(\text{Variety } n) \times g_n)]/\text{StDev}(\text{All samples}) \times g_{\text{total}}, \quad (2)$$

where g_n is the number of samples in group n.

Now if w has a value greater than 1, then the inner variance is greater than the outer variance [64], and as a result this variable is a hindrance to correct discrimination, and so it should definitely be discarded.

The other method used for variable selection is the *Fisher Ratio*, whereby a value is calculated for each variable according to the following formula:

Calculation of the between-group variation:

$$\text{SSB}(a) = \sum_{i=1}^{g} n_i(\bar{y}_i - \bar{y})^2.$$

Calculation of the within-group variation:

$$\text{SSW}(a) = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2.$$

Calculation of the Fisher coefficient F:

$$F = \frac{(1/(g-1))\text{SSB}(a)}{(1/(n-g))\text{SSW}(a)},$$

where g is the number of groups, n_i the number of

elements in i group, $\bar{y}_i$ the mean of group i, $\bar{y}$ the total mean, and y_{ij} is the value of object j in group i.

A value for w is then calculated by taking the inverse of the Fisher ratio, so that the same selection methods can be applied (i.e. the lower the value of w, the better the variable).

The three selection methods are henceforth referred to as Weighted w, Unweighted w and Fisher. The best variables, as judged by these methods, may be found in Table 1.

It may often be found that factors other than that searched for (e.g. olive oil variety) may be having some influence on the data (for example, time of harvesting or region of origin). Although variables containing data so affected may have a value for w less than 1, they may still be having a great influence on the model by causing oils of, say, a similar harvesting date, to cluster together, or at least to be pulled away from their hoped-for varietal clustering, in a PCA scores plot. In order to eliminate this effect, we would wish to discard many variables with a value of w less than 1. As we cannot easily know which variables are affected most in this way, one solution is to start from a minimum number of variables (two or three), selected with a low threshold value of w, and work upwards towards $w=1$, to see at what point the best model is reached.

It could be expected, when trying to identify varieties, that the optimum model will be achieved at or near a value for w which selects variables that contribute to the discrimination of varieties, but does not select any that also contribute a large amount to the discrimination of other factors (such as region or harvesting date). In practice it is found that the ideal threshold ranges from that which selects only the best three or four variables right up to 1, depending on the data and the desired factor (whether variety, region, etc.).

The Excel 5 macros perform tasks such as calculating the value of w for each variable, writing the best n (n is the user's choice) out to another sheet, running PLS, PCR and PCA on the data and graphing the results. The graphs reproduced here have been generated using these macros.

Note that the scores and loadings charts produced by these macros indicate the value for w, calculated using weighted w, for the factors graphed. In addition, the scores charts give a value called w *clustering*, which is the product of the value of w for each of the two factors graphed. This gives an easy way of comparing the clustering of different charts.

4. Results and discussion

4.1. Using PCA scores to discriminate variety

With weighted w, and Fisher selection, PCA was carried out on all the variables, and on the best n for a range of n (see Table 1). Scores and loadings plots were produced. It was found that the first two principal components were usually the only two containing any useful information for variety discrimination. Examining the loadings plot of all the variables (Fig. 1(a)) shows that the best six variables, as judged by weighted w and Fisher, i.e. CS_4, CS_2 and CS_30, CS_21, CS_22 and CS_5, are given a large weighting. The worst variables, CS_17, CS_41, CS_1.1, CS_19, CS_34 and CS_36 tend to lie around the centre, as would be hoped.

It must be remembered that PCA is an unsupervised method, so (unlike PLS) does not know to which variety a sample belongs. It is therefore particularly encouraging that this method has weighted the best variables most favourably.

The scores plot for all variables is shown in Fig. 1(b). This shows that the Dritta oils, used to verify reproducibility, cluster very tightly. From this, it would appear that the data are reliable. It is also clear from this chart that the varieties are not completely separated. However, by the removal of only two varieties, Coratina (14 objects) and Picholine (8 objects), all the varieties may be separated. Repeating PCA on the five remaining varieties yields the chart shown in Fig. 2, in which the varieties are clearly separated from each other.

4.2. Using PLS to discriminate variety

PLS2 [26] was carried out on the data, with 42 objects in the training set, and 38 in the test set. The training set was chosen to be as representative as possible of the regions of origin for each variety, within the constraints of the size of the dataset. The identity of the oils was encoded as a binary vector with a 1 representing the relevant variety [4,65]. Thus the

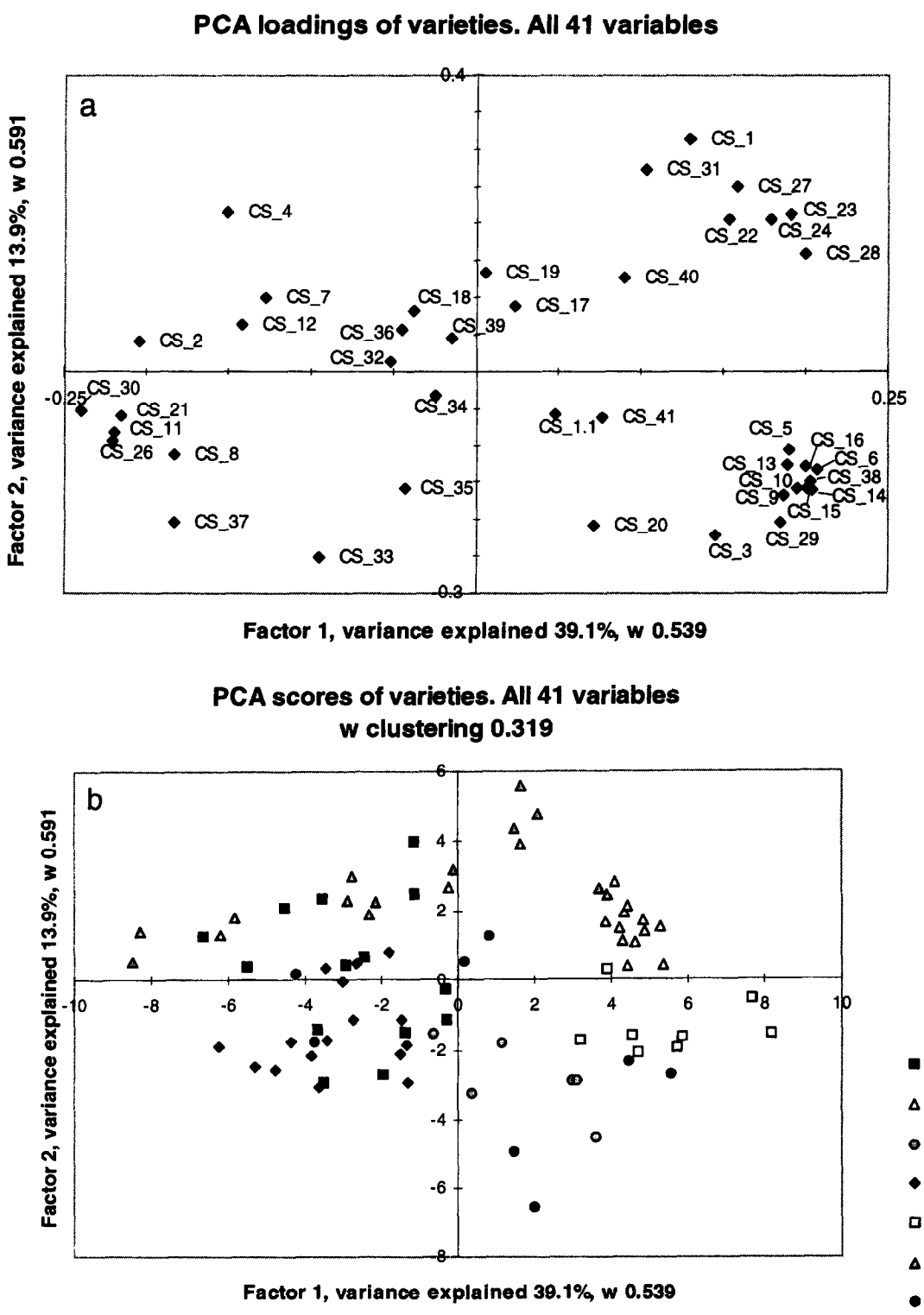

Fig. 1. (a) PCA loadings plot of all variables. (b) PCA scores plot of all 41 variables, factors 1 and 2.

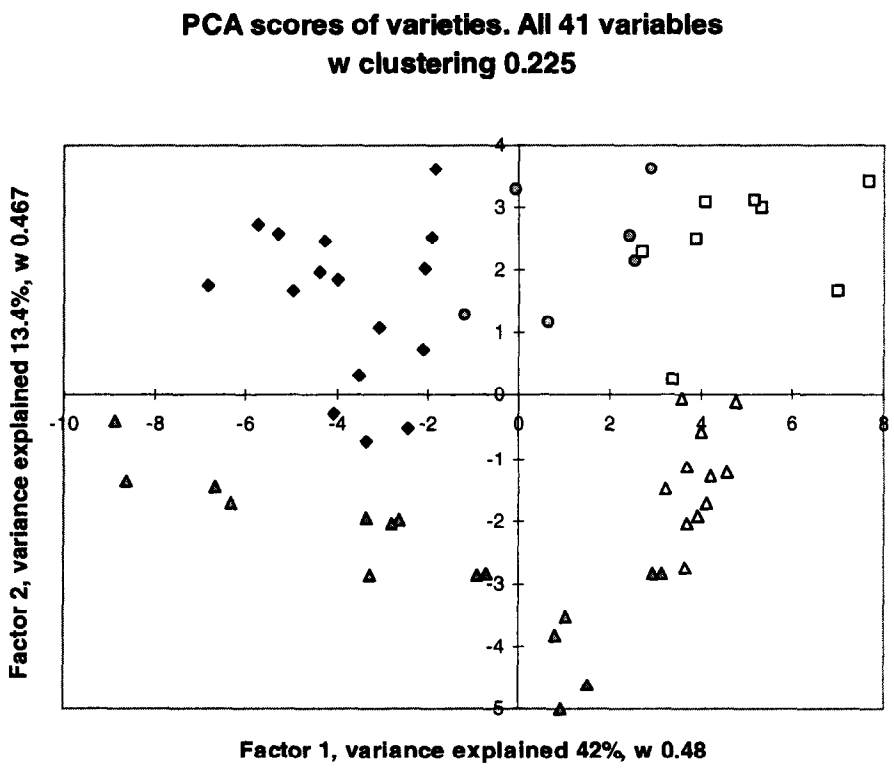

Fig. 2. PCA scores plot of I-77, Grossa di Cassano, Dritta, Moraiolo and Israel oils only, factors 1 and 2.

predicted variety was taken to be that which was given the highest prediction value (i.e. the closest to 1). The optimum number of factors for prediction was chosen by the PLS program as the point at which the RMS error of prediction reaches its first minimum. Comparison of the PLS scores plot (data not shown here) with the PCA scores plot (Fig. 1(b)) shows similar results.

An examination of the loading weights plot (Fig. 3) of the model created using all 41 variables shows very clearly that the best variables according to Fisher and w (Table 1) are given the greatest weighting for this model in the first two factors.

In order to try to improve the clustering, PLS2 predictions were run on the best x variables for all values of x from 41 down to 1, using the Excel macros, with both Fisher and weighted w selection. The best number of variables for prediction were found to be the best 35 and the best 39 for Fisher, and the best 26 and the best 35 for weighted w, both of which correctly identified 68.4% of the samples (Fig. 4(a)). The RMS error of prediction reached a minimum at the best 35 variables for both Fisher and weighted w (Fig. 4(b)). The worst six variables were the same for Fisher and weighted w, so the best 35 models are identical.

PCA and PLS2 scores plots were examined at the points where the best predictions were obtained, and it was found that PCA clusters better than PLS2 scores for many models based on just the best n variables. It was therefore of interest that PCR could actually predict slightly better than PLS (Fig. 5(a)), with unweighted w correctly predicting 73.7% of the oils, and Fisher achieving 71%, both better than the maximum 68.4% achieved by Fisher and weighted w using PLS2 (Fig. 4). The RMSEP (Fig. 5(b)) again shows a low point well before all 41 variables are reached.

By comparison, MLR performs less well than the other methods (data not shown), achieving a best prediction of 65.8%, although it performs at its best with fewer variables. Variable selection appears to be much more crucial for MLR prediction. MLR was

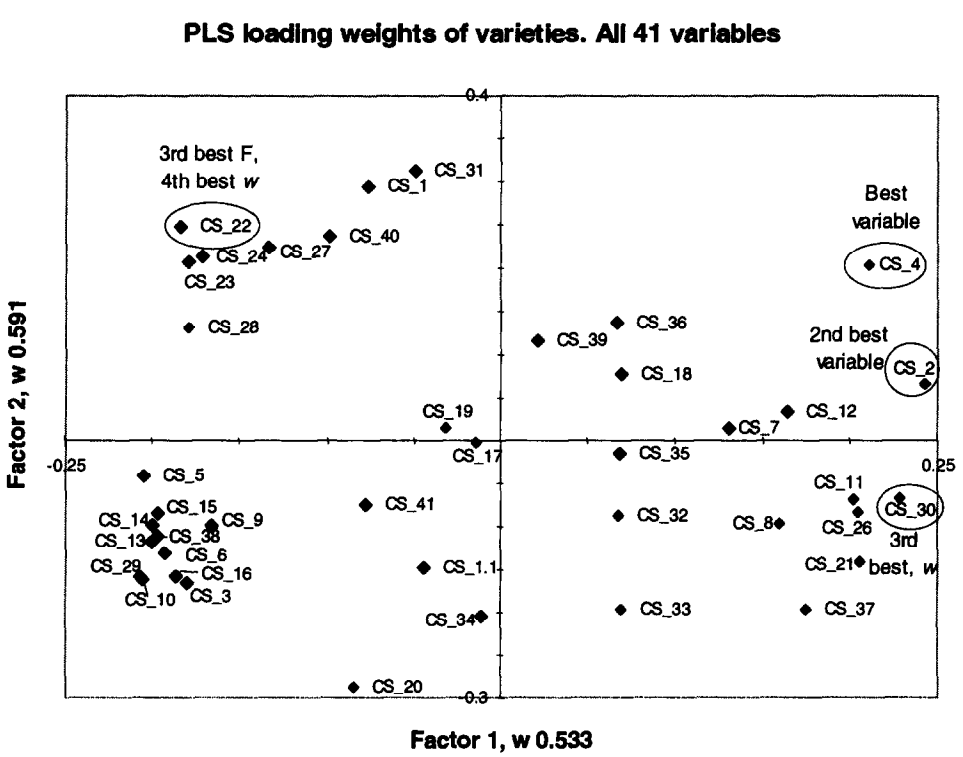

Fig. 3. PLS2 Loading weights plot of varieties, all 41 variables.

always found to be worse than both PLS and PCR and is not further considered here.

4.3. Prediction of specific varieties

Predicting *all* the varieties from the 80 oil samples is clearly not easy, with at best around 70% of the samples being correctly identified. However, within the data there are four varieties with sufficient representation to attempt to predict them alone. These are Coratina (14 objects), I-77 (16), Moraiolo (16) and Dritta (Repro, 12). In addition to these four, it would be hoped that the Israel oil (8), whose variety is unknown, would also be predictable, since it is the only non-Italian oil in the dataset.

The test set and training sets must be carefully chosen in order to obtain the optimum calibration model and hence prediction; this was done by examining the scores plots and selecting for the training set those oils near the edge of the clusters. Half of the 80 oils, including half of the oils in the variety being predicted were reserved for the test set in each case, and the duplicates were kept together. The exception was the Coratina oils, for which three of the seven duplicates (6 from 14 samples) were used in the test set. The success of these predictions is almost complete, and shows much more clearly the advantages of variable selection.

With the predictions from PLS1 and PCR1, a threshold value is set. Values above this threshold are taken to be predictions of the variety in question; those below are taken to be predictions of another variety. This threshold is set to an optimum for the test set for each complete set of results (best 1 to all 41 variables). The charts shown in Fig. 6(a)–(e) are a combination of three sets of results, i.e. Fisher, weighted w and unweighted w selected. Since the optimum threshold is not necessarily the same for each selection method, predictions using all 41 variables do not necessarily correspond in the graphs.

The graphs show very clearly the advantage of variable selection; only the Israel oil achieved its best result using all variables.

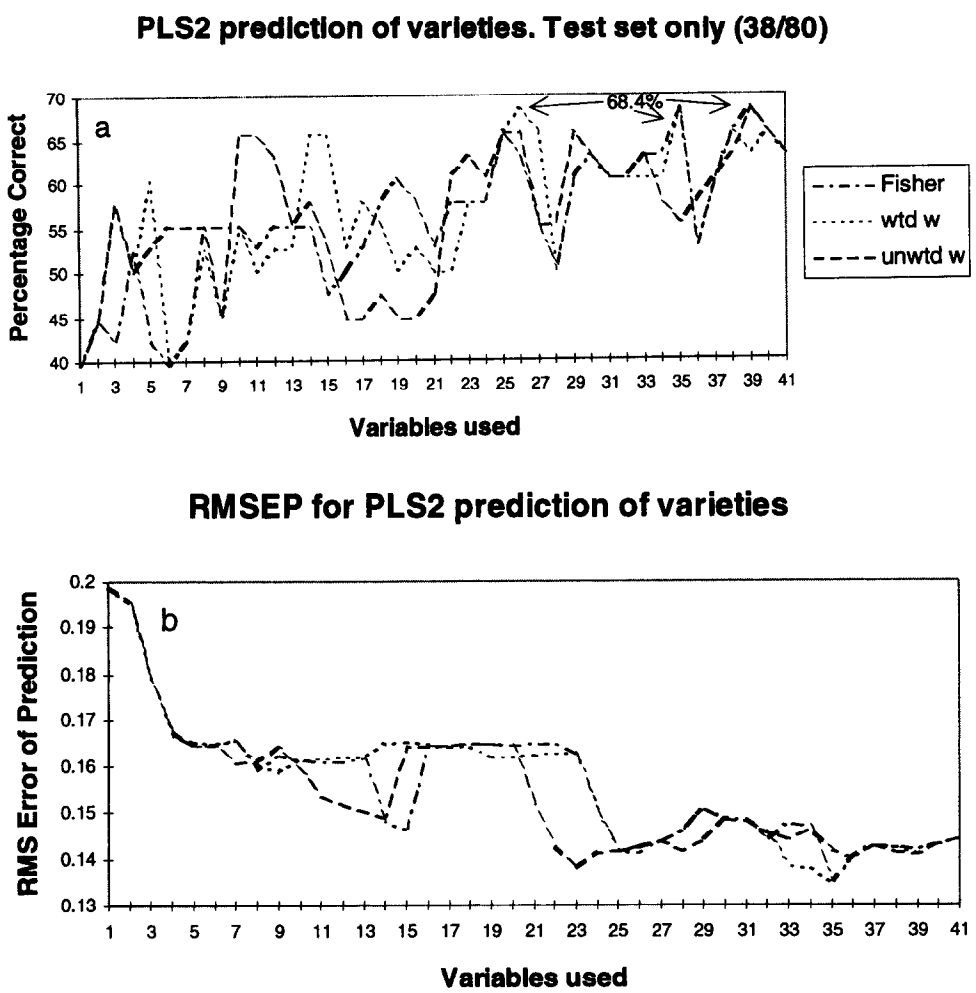

Fig. 4. (a) Percentage of variety samples correctly identified using all the oils, PLS2 and variable selection. (b) RMSEP of PLS2 variety prediction using weighted w, unweighted w and Fisher selection methods.

The Dritta (repro) oils (Fig. 6(a)) predicted easily, despite having only six objects in the training set and six in the test set; this was expected, since all 12 were from the same original sample.

Only the Moraiolo variety (Fig. 6(d)) failed to predict with 100% accuracy in the test set, and then only one sample, a Moraiolo, was incorrectly predicted.

With so few samples from the varieties Grossa di Cassano and Picholine, it was not expected that these would be predictable, as at most four samples could be included in the training set; this was indeed found to be the case (data not shown).

Using this method of prediction, then, only one sample (1.5%) from the 66 Dritta, I-77, Coratina, Moraiolo and Israel samples was not predicted correctly. The optimum number of variables varies from the best six for I-77 (Fig. 6(c)) to all variables for Israel (Fig. 6(e)). This variation in the optimum number of variables is clearly one reason why PLS2 could not perform so well as PLS1.

4.4. Prediction of region

With this set of data, there is sufficient representation from four regions to attempt to make a prediction

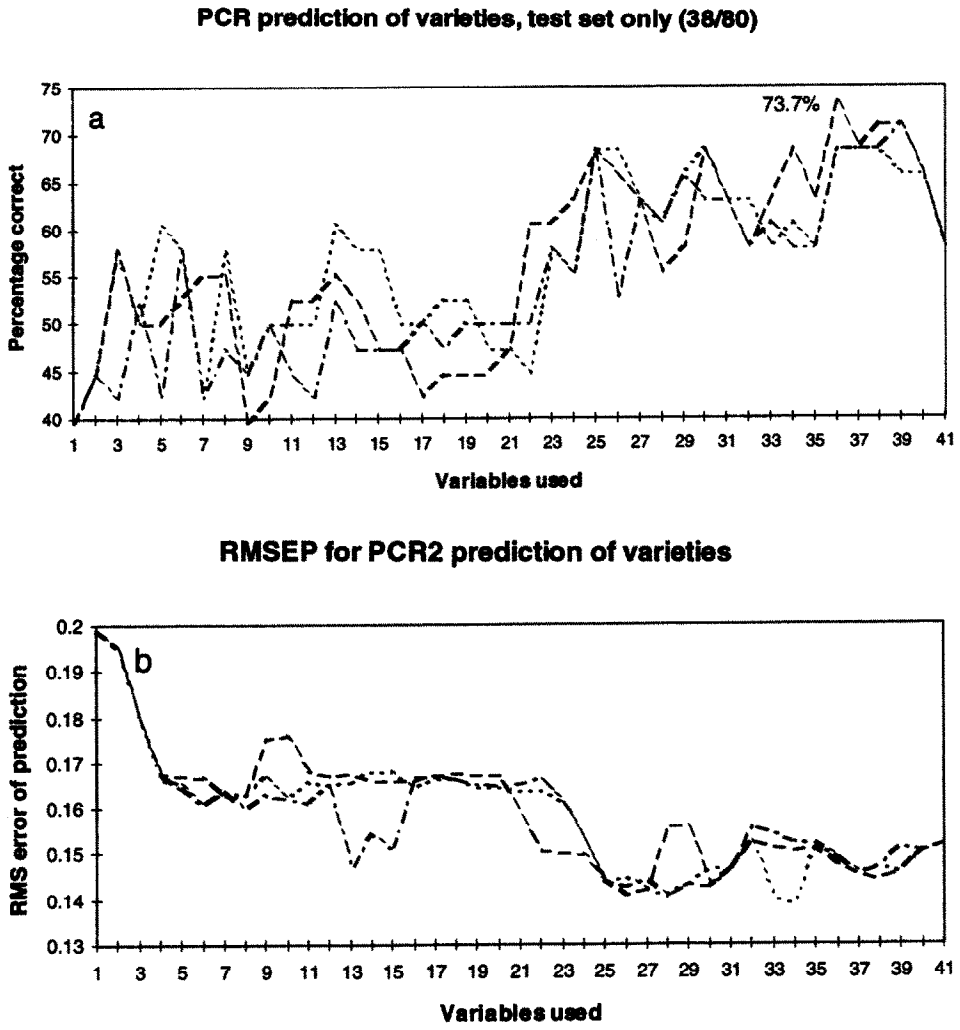

Fig. 5. (a) Percentage of variety samples correctly identified using all the oils, PCR and variable selection. (b) RMSEP of PCR2 variety prediction using weighted w, unweighted w and Fisher selection methods.

of the region of origin of the oil. These regions are: Toscana (Tuscany) (12), Abruzzo (12), Puglia (14) and Israel (8). In addition, the 12 reproducibility (Dritta) oils may be included. Oils from the remaining regions were eliminated from the data, as they contained too few samples to hope to make a prediction.

It is to be expected that region identification would prove much more difficult than variety identification, since a variety is a very definite quantity and a region is no more than an arbitrary line drawn on a map. (Indeed the variance in pyrolysis mass spectra of olive oils seems to be dominated by the variance due to the different varieties, even when considering oils that have been adulterated with other seed oils [4].) Nonetheless, region prediction is probably more important than variety prediction from an economic point of view, as indicated in the introduction, because of the common practice of mislabelling oils to make them appear to be from a more prestigious region.

The regions used here come from geographically widely separated areas: from Italy, Toscana in the north, Abruzzo in the centre and Puglia in the far south, and the far away Israel.

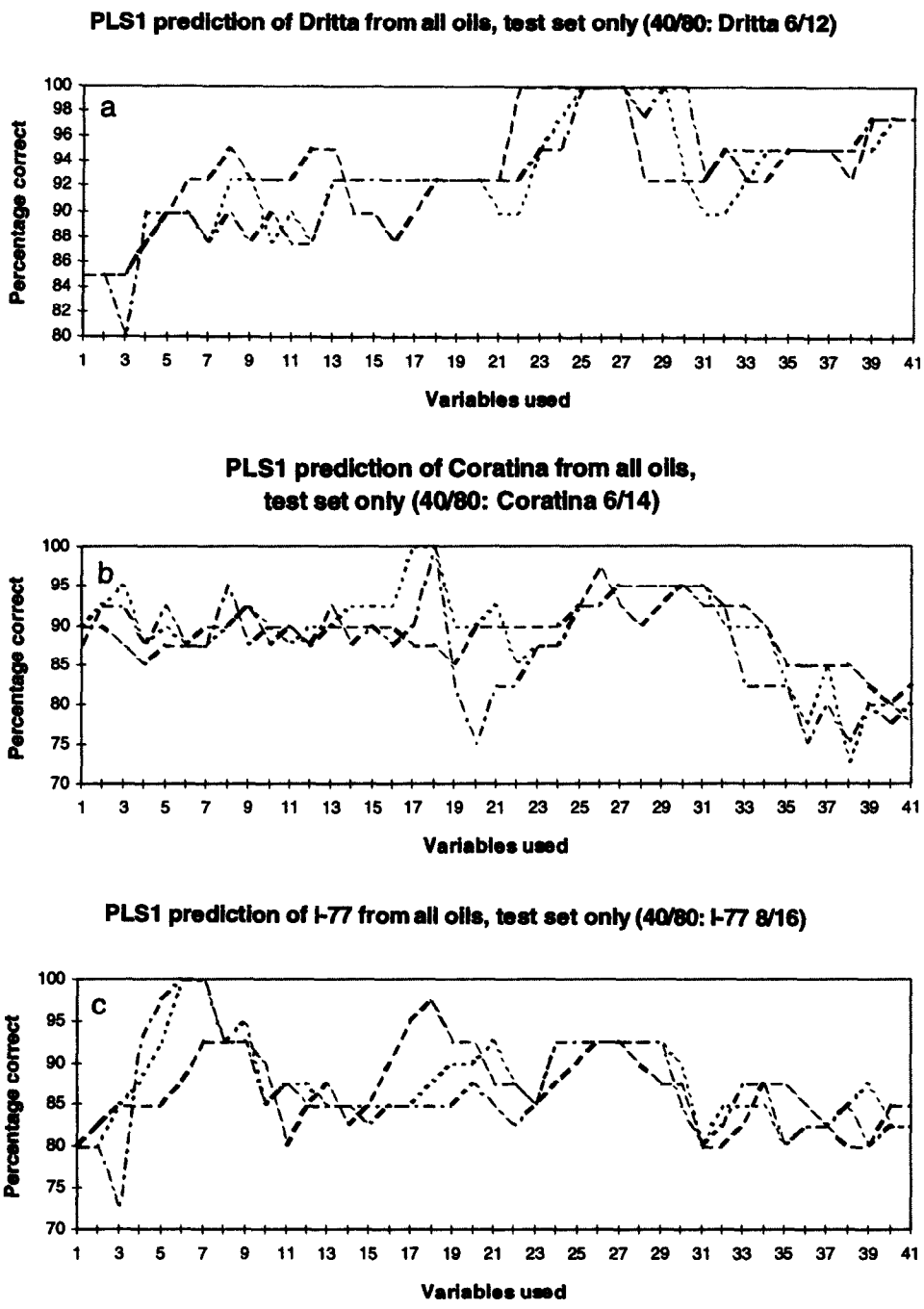

Fig. 6. (a) Dritta PLS1 prediction; all oils correctly predicted with Fisher, weighted w and unweighted w selection, using the best 25 variables. (b) Coratina PLS1 prediction; all oils correctly predicted with Fisher and weighted w selection, using the best 18 variables. (c) I-77 PLS1 prediction; all oils correctly predicted with Fisher and weighted w selection, using the best 6 to 7 variables. (d) Moraiolo PLS1 prediction; all but one oils correctly predicted with Fisher and weighted w selection, using the best 29 variables. (e) Israel PLS1 prediction; all oils correctly predicted with Fisher, weighted w and unweighted w selection, using the best 34 to all 41 variables.

(Continued on next page)

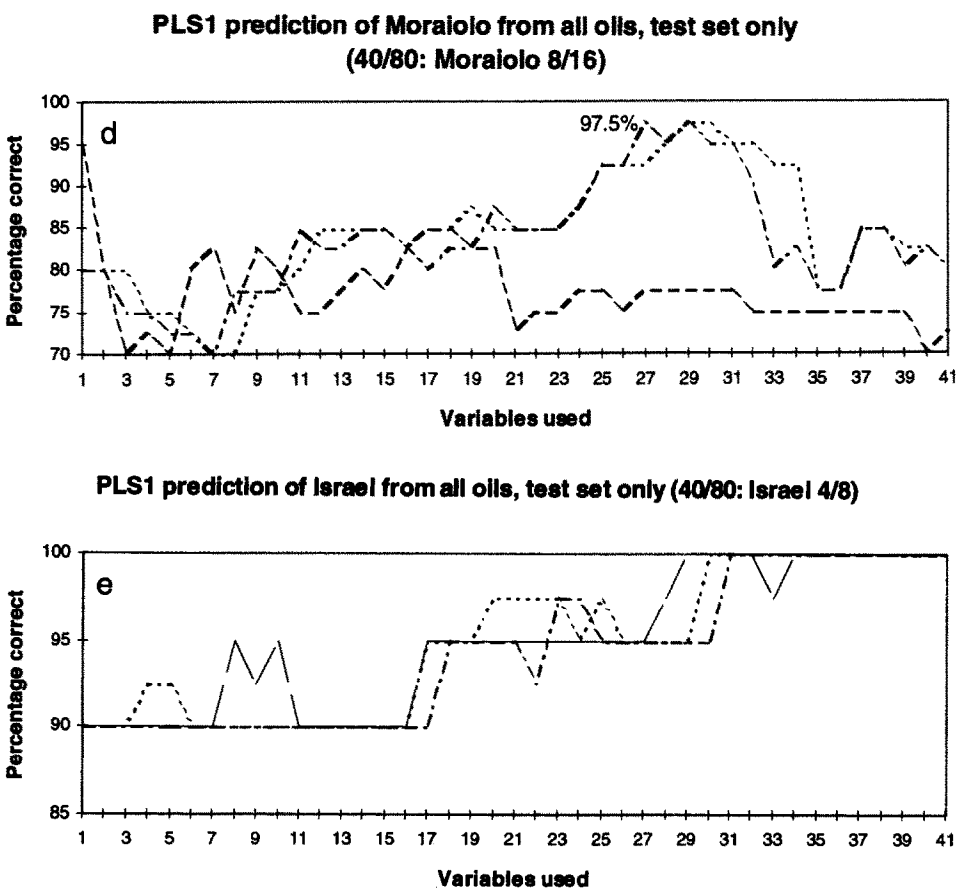

Fig. 6. (continued)

The order of selection of variables for regions is shown in Table 1. Interestingly they are quite different from those of importance in determining variety; for instance CS_11 is of first or second rank for region discrimination, but between the 16th and 20th in importance for variety discrimination.

The successes of the PLS2 (Fig. 7(a)) and PCR2 (Fig. 7(b)) predictions using all the samples from these five regions are comparable with those obtained with the varieties earlier.

The best number of variables is clearly around 18–22, weighted w selected. From the PCA plots of the best 19 variables, weighted w calculated on the factors suggested that factor 1 is the best for discrimination, but factor 4 is the second best (Fig. 8), and factor 2 (not shown) only third. Using all the variables (graph not shown), the first and second factors were found to be best, and the clustering was weaker (w clustering 0.394, compared to 0.378 for the best 19).

Having achieved comparable results with regions as with varieties using PLS2 and PCR2, it may be expected that PLS1 and PCR1 will achieve similar success. The chart in Fig. 9 shows that PLS1 was indeed able to predict 27 from 28 of the samples correctly based upon whether or not the oil was from Tuscany; the only oil incorrectly identified was a Toscana oil.

With only six Toscana oils in the training set, it is perhaps not surprising that it should be a Toscana oil that was incorrect. A larger Toscana sample size would enable more Toscana oils to be put in the training set, and hopefully result in better prediction.

Comparable data (not shown) were obtained from the other regions; 92.9% (all but 2) being the best

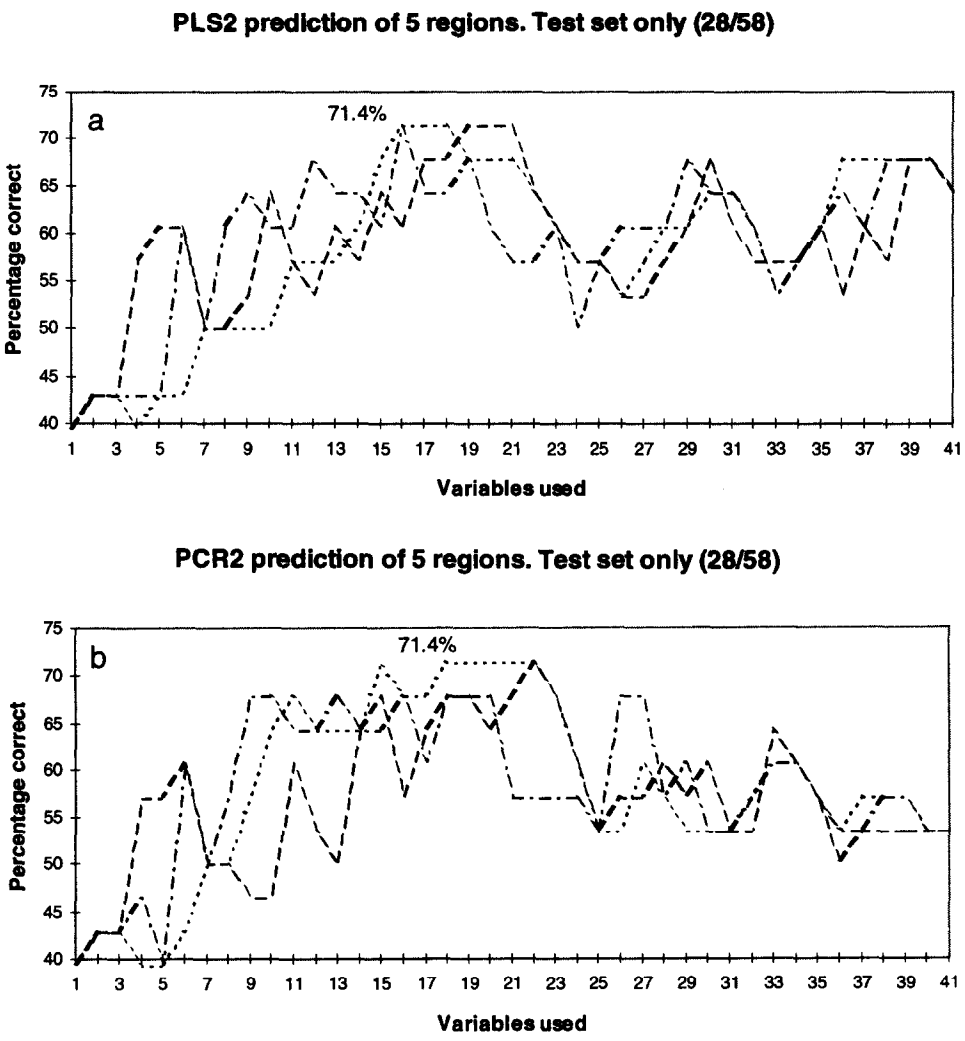

Fig. 7. (a) Percentage of region samples correctly identified using the five largest regions and PLS. (b) Percentage of region samples correctly identified using the five largest regions and PCR.

prediction for region Puglia, 89.7% (all but 3) from Abruzzo, and 100% for Dritta (repro) and Israel.

5. Concluding remarks

From the present results, three main conclusions may be drawn:

- ^{13}C NMR is a valuable tool for the discrimination and classification of olive oils, in terms of both the variety of olive used and the region of cultivation.
- Variable selection greatly improves the predictions of standard statistical techniques on multivariate data.
- Multiple PLS1 models perform far better than does a single PLS2 model (much as is the case with comparable types of artificial neural network [66–69]).

Using these two techniques together, we have been able to predict the variety of all but one test set sample from those where there were sufficient samples for this to be sensible, using only the linear multivariate

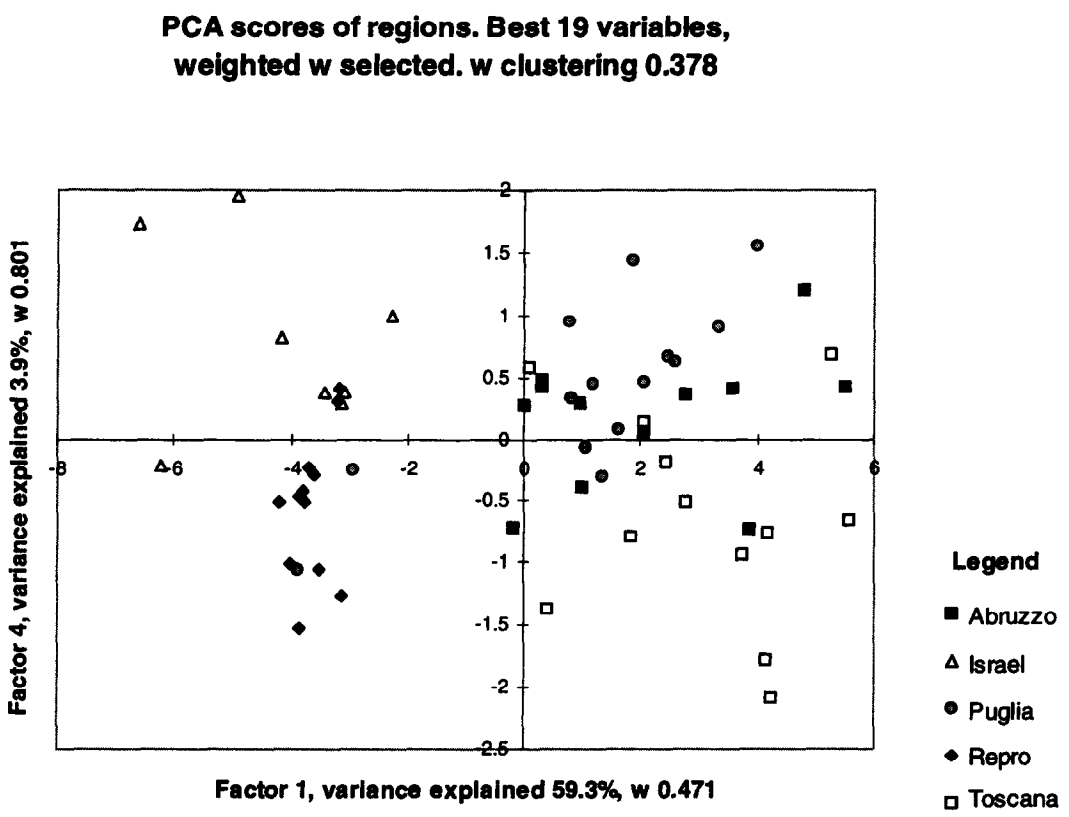

Fig. 8. PCA scores of factors 1 and 4, best 19 variables of largest five regions. Weighted w selected.

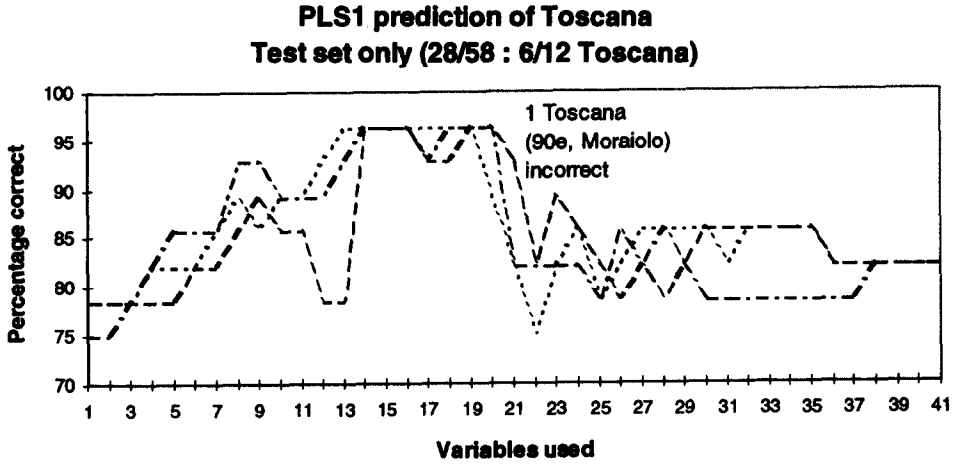

Fig. 9. PLS1 prediction of Toscana region from the five largest regions. Best prediction 96.4%.

method of PLS regression. Finally, we recognise that nonlinear methods such as artificial neural networks (e.g. [70–74]) might have been more successful [4,65,75] (although computational demands with the many models studied meant that they could not be applied here), and that other statistical, neural, sym-

bolic 'expert' and related methods (e.g. [76–83]) remain to be tried. Although the prediction of regions was not quite so successful, we consider that the appropriate way forward is a hierarchical approach in which a model is used to predict the cultivar and based on this submodels are used (given the cultivar) to predict the region. For this, of course, it will be necessary to have larger datasets.

References

[1] A.K. Kiritsakis, American Oil Chemists' Society, 1991.
[2] Anon. in L'Informatore Agrario, Vol. 18, 1994, pp. 41.
[3] D.K. Salunkhe, J.K. Chavan, R.N. Adsule and S.S. Kadam, World Oilseeds: Chemistry, Technology and Utilization, Van Nostrand Reinhold, New York, 1991.
[4] R. Goodacre, D.B. Kell and G. Bianchi, J. Sci. Food Agric., 63 (1993) 297–307.
[5] G.E. Fraser, Am. J. Clin. Nutr., 59 (1994) S1117–S1123.
[6] F.G. Visioli and C. Galli, Life Sci., 55 (1994) 1965–1971.
[7] C. Galli, A. Petroni and F. Visioli, Eur. J. Pharmaceutical Sci., 2 (1994) 67–68.
[8] J.M. Martin-Moreno, W.C. Willett, L. Gorgojo, J.R. Banegas, F. Rodriguez-Artalejo, J.C. Fernandez-Rodriguez, P. Maisonneuve and P. Boyle, Int. J. Cancer, 58 (1994) 774–780.
[9] A. Trichopoulou, K. Katsouyanni, S. Stuver, L. Tzala, C. Gnardellis, E. Rimm and D. Trichopoulos, J. Natl. Cancer Inst. , 87 (1995) 110–116.
[10] A. Trichopoulou, A. Kouris-Blazos, T. Vassilakou, C. Gnardellis, E. Polchronopoulos, M. Venizelos, P. Lagiou, M.L. Wahlqvist and D. Trichopoulos, Am. J. Clin. Nutr., 61(suppl.) (1995) 1346S–1350S.
[11] R. Zamora, J.L. Navarro and F.J. Hidalgo, J. Am. Oil Chem. Soc., 71 (1994) 361–364.
[12] D. Firestone, J.L. Summers, R.J. Reina and W.S. Adams, J. Am. Oil Chem. Soc., 62 (1985) 1558–1562.
[13] D. Firestone, K.L. Carson and R.J. Reina, J. Am. Oil Chem. Soc., 65 (1988) 788–792.
[14] W.N. Aldridge, Toxicology Lett., 64–65 (1992) 59–70.
[15] K. Grob, M. Biedermann, M. Bronz and J.P. Schmid, Z. Lebensm. -Unters. -Forsch., 199 (1994) 191–194.
[16] K. Grob, A.M. Giuffré, U. Leuzzi and B. Mincione, Fat Sci. Technol., 96 (1994) 286–290.
[17] G. Bianchi, A. Tava, G. Vlahov and N. Pozzi, J. Am. Oil Chem. Soc., 71 (1994) 365–369.
[18] R.A. Francelin, F.A.C. Gomide and Lanças, F.M., Chromatographia, 35 (1993) 160–166.
[19] K. Grob, A. Artho and C. Mariani, Fett Wissenschaft Technologie-Fat Sci. Technol., 93 (1991) 494–500.
[20] K. Grob, A.M. Giuffré, M. Biedermann and M. Bronz, Fett Wissenschaft Technologie-Fat Sci. Technol., 96 (1994) 341–345.
[21] R. Goodacre, D.B. Kell and G. Bianchi, J. Sci. Food Agric., 63 (1993) 297–307.
[22] T. Sato, J. Am. Oil Chem. Soc., 71 (1994) 293–298.
[23] I. Schwaiger and F. Vojir, Dtsch. Lebensm. -Rundsch., 90 (1994) 143–146.
[24] Y.W. Lai, E.K. Kemsley and R.H. Wilson, J. Agric. Food Chem., 42 (1994) 1154–1159.
[25] Y.W. Lai, E.K. Kemsley and R.H. Wilson, Food Chem., 53 (1995) 95–98.
[26] H. Martens and T. Næs, Multivariate Calibration, Wiley, New York, 1989.
[27] E. Li-Chan, Trends Food Sci. Technol., 5 (1994) 3–11.
[28] O.M. Kvalheim, D.W. Aksnes, T. Brekke, M.O. Eide, E. Sletten and N. Telnaes, Anal. Chem., 57 (1985) 2858–2864.
[29] T. Brekke, T. Barth, O.M. Kvalheim and E. Sletten, Anal. Chem., 62 (1990) 56–61.
[30] N. Soon, Lipids, 20 (1985) 778.
[31] F.D. Gunstone, Chem. Phys. Lipids, 56 (1990) 195.
[32] K.F. Wollemberg, J. Am. Oil Chem. Soc., 67 (1990) 487.
[33] E. Bengsch, B. Perly, C. Deleuze and A. Valero, J. Magn. Reson., 68 (1986) 1.
[34] U. Horchner and J.H. Kalivas, Anal. Chim. Acta, 311 (1995) 1–13.
[35] U. Horchner and J.H. Kalivas, J. Chemometrics, 9 (1995) 283–308.
[36] J.H. Kalivas, N. Roberts and J.M. Sutter, Anal. Chem., 61 (1989) 2024–2030.
[37] R.G. Brereton and A.K. Elbergali, J. Chemometrics, 8 (1994) 423–437.
[38] R.G. Brereton, Analyst, 120 (1995) 2313–2336.
[39] P.J. Brown, C.H. Spiegelman and M.C. Denham, Philos. Trans. Roy. Soc. London Ser. A, 337 (1991) 311–322.
[40] J.M. Sutter and J.H. Kalivas, Abstracts Of Papers Of the American Chemical Society 203, 24-COMP, 1992.
[41] J.M. Sutter and J.H. Kalivas, Microchem. J. , 47 (1993) 60–66.
[42] P.J. Brown, Measurement, Regression, and Calibration, Oxford Science Publications, Oxford, 1993.
[43] B. Dalmas and W.H. Bannister, Anal. Biochem., 225 (1995) 39–48.
[44] F. Lindgren, P. Geladi, A. Berglund, M. Sjostrom and S. Wold, J. Chemometrics, 9 (1995) 331–342.
[45] C.E. Miller, Chemometrics and Intelligent Laboratory Systems, 30 (1995) 11–22.
[46] L. Norgaard, Talanta, 42 (1995) 1305–1324.
[47] N. Sreerama and R.W. Woody, J. Mol. Biol., 242 (1994) 497–507.
[48] Y.L. Xie, Y.Z. Liang and R.Q. Yu, Anal. Chim. Acta, 272 (1993) 61–72.
[49] D. Rogers and A.J. Hopfinger, J. Chem. Information and Comput. Sci., 34 (1994) 854–866.
[50] D.B. Kell and B. Sonnleitner, Trends Biotechnol., 13 (1995) 481–492.
[51] D. Jouan-Rimbaud, D.L. Massart, R. Leardi and O.E. de Noord, Anal. Chem., 67 (1995) 4295–4301.
[52] M. Baroni, S. Clementi, G. Cruciani, G. Costantino, D. Riganelli and E. Oberrauch, J. Chemometrics, 6 (1992) 347–356.

[53] J.H. Wikel and E.R. Dow, Bioorganic and Medicinal Chem. Lett., 3 (1993) 645–651.
[54] G. Cruciani and K.A. Watson, J. Medicinal Chem., 37 (1994) 2589–2601.
[55] H. Kubinyi, Quantitative Structure-Activity Relationships, 13 (1994) 393–401.
[56] H. Kubinyi, Quantitative Structure-Activity Relationships, 13 (1994) 285–294.
[57] J.M. Sutter, S.L. Dixon and P.C. Jurs, J. Chem. Information and Computer Sci., 35 (1995) 77–84.
[58] H. Kubinyi, J. Chemometrics, 10 (1996) 119–133.
[59] U. Norinder, J. Chemometrics, 10 (1996) 95–105.
[60] M.B. Seasholtz and B. Kowalski, Anal. Chim. Acta, 277 (1993) 165–177.
[61] O.E. de Noord, Chemometrics and Intelligent Laboratory Systems, 23 (1994) 65–70.
[62] D.J. Livingstone and D.T. Manallack, J. Med. Chem., 36 (1993) 1295–1297.
[63] C. Chatfield, J. R. Statist. Soc., 158 (1995) 419–466.
[64] W. Eshuis, P.G. Kistemakker and H.L.C. Meuzelaar, in C.E.R. Jones and C.A. Cramers (Eds.), Elsevier, Amsterdam, 1977, pp. 151–156.
[65] R. Goodacre, D.B. Kell and G. Bianchi, Nature, 359 (1992) 594.
[66] T. Hrycej, Modular Learning in Neural Networks, Wiley, New York, 1992.
[67] M.I. Jordan, J. Math. Psychol., 36 (1992) 396–425.
[68] R. Goodacre, M.J. Neal and D.B. Kell, Anal. Chem., 66 (1994) 1070–1085.
[69] P. Mendes and D.B. Kell, Biosystems, 38 (1996) 15–28.
[70] D.E. Rumelhart, J.L. McClelland and the PDP Research Group, Parallel Distributed Processing: Experiments in the Microstructure of Cognition, MIT Press, Cambridge, MA, 1986.
[71] P.J. Werbos, The Roots of Back-Propagation: From Ordered Derivatives to Neural Networks and Political Forecasting, Wiley, Chichester, 1993.
[72] S.S. Haykin, Neural networks: A comprehensive foundation, Macmillan, New York, 1994.
[73] C.M. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 1995.
[74] B.D. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, Cambridge, 1996.
[75] J. Zupan, M. Novic, X.Z. Li and J. Gasteiger, Anal. Chim. Acta, 292 (1994) 219–234.
[76] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, Classification and regression trees, Wadsworth, Belmont, CA, 1984.
[77] C.W. Therrien, Decision Estimation and Classification: An Introduction to Pattern Recognition and Related Topics, Wiley, New York, 1989.
[78] E. Rich and K. Knight, Artificial Intelligence, McGraw Hill, New York, 1991.
[79] S.H. Weiss and C.A. Kulikowski, Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems, Morgan Kaufmann Publishers, San Mateo, CA, 1991.
[80] J.R. Koza, Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press, Cambridge, MA, 1992.
[81] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.
[82] D. Michie, D.J. Spiegelhalter and C.C. Taylor, in J. Campbell (Ed.), Ellis Horwood Series in Artificial Intelligence, Ellis Horwood, Chichester, UK, 1994.
[83] J.R. Koza, Genetic Programming II: Automatic Discovery of Reusable Programs, MIT Press, Cambridge, MA, 1994.