Turk J Chem
28 (2004) , 497 – 514.

Genetic Multivariate Calibration Methods for Near Infrared (NIR) Spectroscopic Determination of Complex Mixtures

Durmuş ÖZDEMİR and Betül ÖZTÜRK
İzmir Institute of Technology, Faculty of Science, Department of Chemistry,
Gülbahçe Köyü 35437 Urla, İzmir-TURKEY
e-mail: durmusozdemir@iyte.edu.tr

Received 08.10.2003

The simultaneous determination of ternary mixtures of methylene chloride, ethyl acetate, and methanol using near infrared (NIR) spectroscopy and 4 different genetic algorithms based multivariate calibration methods was demonstrated. The 4 genetic multivariate calibration methods are genetic partial least squares (GPLS), genetic regression (GR), genetic classical least squares (GCLS) and genetic inverse least squares (GILS). The sample data set contains the NIR spectra of 63 ternary mixtures and covers the range from 900 to 2000 nm in 2 nm intervals. Of these 63 spectra, 42 were used as the calibration set, and 21 were reserved for the prediction purposes. Several calibration models were built with the 4 genetic algorithm based methods for each component that makes up the mixtures. Overall, the standard error of calibration (SEC) and the standard error of prediction (SEP) were in the range of 0.22 to 2.5 (% by volume (v/v)) for all the 4 methods. A comparison of genetic algorithm selected wavelengths for each component and for each method was also included.

Key Words: Near infrared spectroscopy, Multivariate calibration, Genetic algorithms, Genetic Regression, Partial Least Squares, Classical Least Squares, Inverse Least Squares.

Introduction

Near infrared (NIR) spectroscopy[1,2] has become a popular method for simultaneous chemical analysis and is being studied extensively in a number of different fields such as process monitoring[3], biotechnology[4,5], and the pharmaceutical and food industries[6,7] because of the potential for on-line, rapid, nondestructive and noninvasive instrumentation. The NIR portion of the electromagnetic spectrum covers the range from 780 nm to 2500 nm and most of the absorption bands observed in this region are due to overtones and combinations of the fundamental mid-IR molecular vibration bands. Although all the fundamental vibration modes can have overtones, the most commonly observed bands arise from the C–H, O–H, and N–H bonds in the molecules.

Modern spectroscopic instruments are so fast that they can produce hundreds of spectra in a few minutes for a given sample that contains multiple components. Unfortunately, univariate calibration methods are not suitable for this type of data, as they require an interference free system. Multivariate calibration

deals with data containing instrument responses measured on multiple wavelengths for a sample that usually contains more than one component. In recent years, advances in chemometrics and computers have lead to the development of several multivariate calibration methods[4−7] for the analysis of complex chemical mixtures.

Genetic regression (GR)[8−12] is a calibration technique that optimizes linear regression models using a genetic algorithm (GA) and it has been applied to a number of multi-instrument calibration and wavelength selection problems. GAs are non local search and optimization methods that are based upon the principles of natural selection[13−15]. For a given full spectrum of data, GR selects an optimum linear combination of wavelengths and simple mathematical operators to build a linear calibration model using the simple least squares method.

Classical least squares (CLS) extends the classical Beer's Law model in which the absorbance at each wavelength is directly proportional to the component concentrations. Inverse least squares (ILS) is based on the inverse Beer's Law where concentrations of an analyte are modeled as a function of absorbance measurements. Genetic classical least squares (GCLS) and genetic inverse least squares (GILS) are modified versions of the original CLS and ILS[16,17] methods in which a small set of wavelengths is selected from a full spectral data matrix and evolved to an optimum solution using a genetic algorithm.

Partial least squares (PLS) is a soft modeling technique in which the data are decomposed into new variables that are linear combinations of the original data. These new variables are named as principal components or factors and, therefore, PLS are often called as a factor method. Genetic partial least squares (GPLS) is a modification of the NIPALS algorithm in which a genetic algorithm is used to select a set of wavelengths while determining the first factor. CLS, ILS, and PLS have been well described by Wold, Haaland and Kowalski and co-workers[16−19].

In this study, 4 different genetic algorithm based calibration methods GR, GPLS, GCLS and GILS were tested with the aim of establishing calibration models that have a high predictive capacity for the simultaneous determination of methylene chloride, ethyl acetate, and methanol in their binary and ternary mixtures using the NIR spectroscopic technique. The genetic algorithms used in CLS and PLS were the same but they had some differences in GR and GILS. For this reason, a comparison of selected wavelengths not only for each method but also for each component is given.

Theory

Genetic regression

Genetic algorithms (GAs) are global search and optimization methods based upon the principles of natural evolution and selection as developed by Darwin. Computationally, the implementation of a typical GA is quite simple and consists of 5 basic steps including initialization of a gene population, evaluation of the population, selection of the parent genes for breeding and mating, crossover and mutation, and replacing parents with their offspring. These steps have taken their names from the biological foundation of the algorithm. Genetic regression (GR) is an implementation of a GA for selecting wavelengths and mathematical operators to build linear calibration models. GR is a hybrid calibration method between univariate and multivariate calibration techniques that optimizes simple linear regression models through an evolving selection of wavelengths and simple mathematical operators (+, -, *, /). GR follows the same basic initialize/ breed/mutate/evaluate algorithm as other GA's but differs in the way it encodes genes. A 'gene' is a potential

solution to a given problem and the exact form may vary from application to application. Here, the term 'gene' is used to describe the collection of instrument response pairs combined with the above mentioned operators. These pairs, called 'base pairs', are then combined with an addition operator to produce a score, which relates the instrument response to the component concentration. The term 'population' is used to describe the collection of individual genes in the current generation.

In the initialization step, the first generation of genes is created randomly with a fixed population size. Although random initialization helps to minimize bias and maximize the number of possible recombinations, GR is designed to select initial genes in a somewhat biased random fashion in order to start with genes better suited to the problem than those that would be randomly selected. Biasing is done with a correlation coefficient by plotting the scores of initial genes against the component concentrations. The size of the gene pool is a user defined even number in order to allow breeding of each gene in the population. It is important to note that the larger the population size, the longer the computation time. The number of base pairs in a gene is determined randomly between a fixed low limit and high limit. The lower limit was set to 2 in order to allow single point crossover whereas the higher limit was set to eliminate overfitting problems and reduce the computation time. Once the initial gene population is created, the next step is to evaluate and rank the genes using a fitness function, which is the inverse of the standard error of calibration (SEC) given as:

$$SEC = \sqrt{\frac{\sum_{i=1}^{m} (\hat{c}_i - c_i)^2}{m-2}} \tag{1}$$

where m is the number of calibration samples, $\hat{c}_i$is the predicted concentration and $c_i$is the actual concentration. In order to test the performance of each gene, cross validation is applied to the samples in the calibration set where each model was constructed with $m-1$ number of samples and the left out sample spectrum was used to validate the model. This process continues until each spectrum is left out once in the calibration set.

The third step is where the basic principle of natural evolution is put to work for GR. This step involves the selection of the parent genes from the current population for breeding using a roulette wheel selection method according to their fitness values. The goal is to give a higher chance to those genes with high fitness so that only the best performing members of the population will survive in the long run and will be able to pass their information to the next generations. Because of the random nature of the roulette wheel selection method, however, genes with low fitness values will also have some chance to be selected. In addition, there will be genes that are selected multiple times and some genes will not be selected at all and will be thrown out of the gene pool. After the selection procedure is completed, the selected genes are allowed to mate top-down without ranking whereby the first selected gene mates with the second gene and the third one with the fourth one and so on as illustrated in the following example:

Parents

$$S_1 = (A_{347} * A_{251})\# + (A_{379} + A_{218})(2) \tag{2}$$

$$S_2 = (A_{225} * A_{478})\# + (A_{343}/A_{250}) + (A_{451} - A_{358}) + (A_{231} - A_{458)(3)} \tag{3}$$

The points where the genes are cut for mating are indicated by #.

Offspring

$$S_3 = (A_{347} * A_{251}) + (A_{343}/A_{250}) + (A_{451} - A_{358}) + (A_{231} - A_{458)(4)} \tag{4}$$

$$S_4 = (A_{225} * A_{478}) + (A_{379} + A_{218})(5) \tag{5}$$

Here the first part of S_1 is combined with the second part of S_2 to give S_3; likewise the second part of S_1 combined with the first part of S_2 to give S_4. This process is called the singe point crossover and is the one used in GR. The single point crossover will not provide different offspring if both parent genes are identical, which may happen in the roulette wheel selection, and are broken at the same point. Also note that mating can increase or decrease the number of base pairs in the offspring genes. After crossover, the parent genes are replaced by their offspring and the offspring are evaluated. The ranking process is based on their fitness values following the evaluation step. Then the selection for breeding/mating starts all over again. This is repeated until a predefined number of iterations is reached.

In the end, the gene with the lowest SEC (highest fitness) is selected for model building, which is done by simple least squares. This model is used to predict the concentrations of component being analyzed in the validation (test) sets. The success of the model in the prediction of the validation sets is evaluated using standard error of prediction (SEP), which is similar to SEC except that the degrees of freedom is m not $m - 2$. Notice that the validation set is not used during the model building step at all. Because the random processes are heavily involved in GR, as in all the GAs, the program was set to run 25 times for each component in a given multi-component mixture during the course of this study. The best run (i.e. the one generating the lowest SEC for the calibration set and at the same time producing SEPs for validation sets that are in the same range as the SEC) was subsequently selected for evaluation and further analysis. The termination of the algorithm can be done in many ways. The easiest way is to set a predefined iteration number for the number of breeding/mating cycles.

GR has some major advantages over classical univariate and multivariate calibration methods. It is a hybrid calibration method that uses the full spectral information and reduces it to a single score upon which simple calibration models are built. First of all, it is as simple as univariate calibration in terms of the mathematics involved in the model building and prediction steps, but at the same time it has the advantages of the multivariate calibration methods since it uses the full spectrum to extract genetic scores. It automatically corrects baseline fluctuations with the use of simple mathematical operators while forming the base pairs.

Genetic classical least squares

The classical least squares (CLS) method extends the classical Beer's Law model in which the absorbance at each wavelength is directly proportional to the component concentrations. Model errors are assumed to be in the measurement of the instrument responses as it was the case in the classical univariate method. In matrix notation, the CLS model for m calibration samples containing l chemical components whose spectra contains n wavelengths is described as:

$$\mathbf{A} = \mathbf{CK} + \mathbf{E_A} \tag{6}$$

where $\mathbf{A}$ is the mxn matrix of the calibration spectra, $\mathbf{C}$ is the mxl matrix of the component concentrations, $\mathbf{K}$ is the lxn matrix of absorptivity-pathlength constants and $\mathbf{E}_A$ is the mxn matrix of the spectral errors or residuals not fit by the model. Here the $\mathbf{K}$ matrix represents the first order estimates of the pure component spectra at unit concentration and unit pathlength. The least-squares method can be used to estimate the

K matrix. The least squares estimate of **K** is defined as

$$\hat{\mathbf{K}} = (\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'\mathbf{A} \tag{7}$$

Once the estimated $\hat{\mathbf{K}}$matrix is obtained, the concentrations of an unknown sample can be predicted from its spectrum by:

$$\hat{\mathbf{c}} = (\hat{\mathbf{K}}\hat{\mathbf{K}}')^{-1}\hat{\mathbf{K}}\mathbf{a} \tag{8}$$

where **a** is the spectrum of the unknown sample and $\hat{\mathbf{c}}$is the vector of the predicted component concentrations. Genetic classical least squares (GCLS) is a modified version of the original CLS method in which a small set of wavelengths is selected from a full spectral data using a genetic algorithm. The algorithm used to select the optimum number of wavelengths in GCLS is quite similar to the GR algorithm, but differs in the way it encodes the gene. In GCLS, the term 'gene' describes a vector whose elements are randomly selected wavelengths and there are no base pairs as explained in GR. The size of the vector is also determined in a random fashion with an upper limit to reduce computation time.

In the initialization step, an even number of genes are formed from full a spectral data matrix and each gene is used to form a CLS model. These models are then evaluated and ranked using the fitness function described in GR. The roulette wheel method is then used to select the gene population for breeding. After the selection procedure is completed, the selected genes are allowed to mate top-down without ranking whereby the first gene mates with the second gene and the third gene with the fourth gene and so on as described above, but with one difference. Since the genes used in GCLS are only vectors of wavelengths and contain no base pairs as described in GR, for each gene a random number is generated between 1 and the length of the gene, and the single point crossover process is performed using this number. After crossover, the parent genes are replaced by their offspring and the offspring are evaluated. The ranking process is based on their fitness values and follows the evaluation step. Then the selection for breeding/mating starts all over again. This is repeated until a predefined number of iterations is reached. During each iteration the best gene with the lowest SEC is stored in order to compare it with the best gene of the next generation. If the next generation produces a better gene then it replaces by the older one; otherwise the old one is kept for further iterations. At the end, the gene with the lowest SEC is selected for model building. This model is used to predict the concentrations of component being analyzed in the validation (test) sets as described in GR.

Genetic inverse least squares

The major drawback of CLS is that all of the interfering species must be known and their concentrations included in the model. This need can be eliminated by using the inverse least squares (ILS) method which uses the inverse of Beer's Law. In the ILS method, concentrations of an analyte are modeled as a function of absorbance measurements. Because modern spectroscopic instruments are very stable and provide excellent signal-to-noise (S/N) ratios, it is believed that the majority of errors lie in the reference values of the calibration sample, not in the measurement of their spectra. The ILS model for m calibration samples with n wavelengths for each spectrum is described by

$$\mathbf{C} = \mathbf{AP} + \mathbf{E_C} \tag{9}$$

where **C** and **A** are the same as in CLS, **P** is the nxl matrix of the unknown calibration coefficients relating l component concentrations to the spectral intensities and $\mathbf{E}_C$ is the mxl matrix of errors in the

concentrations not fit by the model. In the calibration step, ILS minimizes the squared sum of the residuals in the concentrations. The greatest advantage of ILS is that Equation 9 can be reduced for the analysis of a single component at a time since analysis is based on an ILS model that is invariant with respect to the number of chemical components included in the analysis. The reduced model is given as:

$$\mathbf{c} = \mathbf{Ap} + \mathbf{e_c} \tag{10}$$

where $\mathbf{c}$ is the $mx1$ vector of concentrations for the analyte that is being analyzed, $\mathbf{p}$ is $nx1$ vector of calibration coefficients and $\mathbf{e}_c$is the $mx1$ vector of concentration residuals not fit by the model. During the calibration step, the least-squares estimate of $\mathbf{p}$ is

$$\hat{\mathbf{p}} = (\mathbf{A}'\mathbf{A})^{-\mathbf{1}}\mathbf{A}' \cdot \mathbf{c} \tag{11}$$

where $\hat{\mathbf{p}}$is the estimated calibration coefficients. Once $\hat{\mathbf{p}}$is calculated, the concentration of the analyte of interest can be predicted with the equation

$$\hat{c} = \mathbf{a}' \cdot \hat{\mathbf{p}} \tag{12}$$

where $\hat{c}$is the scalar estimated concentration and $\mathbf{a}$ is the spectrum of the unknown sample. The ability to predict one component at a time without knowing the concentrations of interfering species has made ILS one of the most frequently used calibration methods. However, the identity of the interfering species still needs to be known to prepare a good calibration sample set.

The major disadvantage of ILS can be seen in Equation 11 where the matrix, which must be inverted, has dimensions equal to the number of wavelengths in the spectrum and this number needs to be equal to or smaller than the number of calibration samples. This is a severe restriction since the number of wavelengths in a spectrum will generally be greater than the number of calibration samples, and the selection of wavelengths that provide the best fit for the model is not a trivial process. Several wavelength selection strategies, such as stepwise wavelength selection and all possible combination searches, are available to build an ILS model that fits the data best. Here we used the same genetic algorithm described in GCLS to build genetic inverse least squares (GILS) models with one difference. This difference is in the way the mating and single point crossover operations are carried out. Because the number of wavelengths is restricted in response matrix $\mathbf{A}$ in ILS, the size of the largest gene is restricted to one less than the number of calibration samples in the concentration vector. However, if the single point crossover is set to take place at any point of a gene, then the mating step could produce new genes that have a larger number of wavelengths than the number of calibration samples even though all the genes in the initial gene pool were set to have a smaller number of wavelengths than the size of the concentration vector. In order to avoid this problem, the crossover operation is only performed in the middle of each gene in GILS so that the new generations will never have larger sizes than the number of calibration samples. The rest of the algorithm is the same as the one used in GCLS.

Genetic partial least squares

Partial least squares (PLS) is a soft modeling technique in which the data are decomposed into new variables that are linear combinations of the original data. These new variables are named as principal components or factors. The way in which the new variables are created can be visualized for a 2–dimensional system. If the instrument responses for a set of m samples at 2 wavelengths ($n = 2$) are plotted against each other, a

new axis is formed in the direction that represents maximum variability of the data. This new axis is called the first principal component or first eigenvector. If all of the samples fall on this new axis, then all of the variations can be described using only one eigenvector[4]. Otherwise a second eigenvector can be found that is perpendicular or orthogonal to the first eigenvector. The second one describes the maximum amount of residuals, not fit by the first one, in the data set and so on. If more than 2 wavelengths are included in instrument response matrix, the plotting space becomes multidimensional and several eigenvectors can be found, each one successively accounting for the maximum possible amount of remaining variability and each orthogonal to the others. The general model for PLS can be described as

$$\mathbf{A} = \mathbf{TB} + \mathbf{E_A} \tag{13}$$

where **A** is the same as in GCLS and GILS, **B** is a hxn matrix of basis vectors or loading spectra, and **T** is an mxh matrix of intensities or scores in the new coordinate system defined by the h loading vectors. $\mathbf{E}_A$ is now the mxn matrix of spectral residuals not fit by the factor model. The number of basis vectors, h, to represent original calibration spectra is determined by an algorithm during the calibration step. The **T** and **B** matrices are calculated in a stepwise manner (one vector at a time) until the desired model has been obtained using a modified version of the NIPALS[16] (nonlinear iterative partial least squares) algorithm. The process of determining the optimal number of PLS factors may vary from algorithm to algorithm. The cross-validation approach, which is used in this study, is one of the methods for this.

Genetic partial least squares (GPLS) is a modification of the above PLS algorithm in which a genetic algorithm is used to select a set of wavelengths from an original full spectrum data matrix to form the PLS calibration models. In GPLS, a set of randomly selected wavelengths is used to form the gene as described in GCLS and GILS and then this gene is used to construct a new data matrix (A). With this smaller data matrix, a PLS model is generated based on the first factor calculated using the NIPALS algorithm briefly described above. A simple least–squares method is used to regress self predicted calibration concentrations against the actual values and a correlation coefficient (r^2) is calculated for the selected gene. This correlation coefficient is then compared with a predefined r^2 value. If the calculated r^2 is greater than the predefined value, then this gene is selected for further processing, otherwise it is discarded and a new one is selected. Thus, in the initial step of GPLS, a gene pool is generated based on the above procedure. Once this gene pool is generated, the genes are sorted based on their fitness values and allowed to breed using single point crossover as in GCLS. Once the breeding and mating step is over the old generation is replaced by the new one and new PLS models are generated based on the first factor. The process described up to this point is repeated a user defined number of iterations and in each iteration the gene that has the smallest SEC is reserved in order to compare it with the next generation's best gene. At the end of the genetic algorithm step, the gene with the lowest SEC is selected to build the final PLS calibration model. Note that, up to this point several PLS models have been generated, but based only on the first factor. Now with the final gene a full PLS model is generated and number of PLS factors are determined by the cross–validation approach and SEC is calculated for the calibration set. After the model is generated, this model is then tested with an independent validation set and SEP is calculated. As can be seen, the genetic algorithm used in GCLS, GILS and GPLS is not only used to select a set of wavelengths but also put them in competition through an evolving algorithm.

Experimental

Sample preparation and instrumentation

The samples used in this study were binary and ternary mixtures of methylene chloride, ethyl acetate, and methanol. All solvents were HPLC grade obtained from J.T. Baker and were kept over molecular sieves to remove trace water. Three sets of sample mixture were prepared each having the same percentage composition. Table 1 illustrates the concentration profiles in the first sample set. Two different calibration set designs were carried out in order to develop models. In one case, the first sample set was used to prepare the calibration set and the third set was used to prepare the validation set. In another design, the first and second sample sets were used to prepare the calibration model whereas the third set was reserved for the validation set.

Table 1. Concentration profiles for the 21 samples in the calibration and validation sets.

Methylene Chloride (v/v %)	Ethyl Acetate (v/v %)	Methanol (v/v %)
0	100	0
0	80	20
0	60	40
0	40	60
0	20	80
0	0	100
20	80	0
20	60	20
20	40	40
20	20	60
20	0	80
40	60	0
40	40	20
40	20	40
40	0	60
60	40	0
60	20	20
60	0	40
80	20	0
80	0	20
100	0	0

Near infrared absorbance spectra of mixtures were measured between 900 and 2000 nm at 2 nm resolution on a dispersive Shimadzu 3100 UV-Vis-NIR dual beam spectrophotometer using a PbS detector. Quartz cuvettes having a pathlength of 10 mm were used and the blank for all measurements was carbon tetrachloride, which has no appreciable absorbance in the NIR range used in the data processing.

Data Processing and Software

The spectra were transferred to a separate PC after collection on the instrument and MS Excel (Microsoft Office 97, Microsoft Corporation) was used to prepare text files that are required for the methods used in this study. The 4 new genetic algorithms based multivariate calibration methods (GPLS, GR, GCLS, and

GILS) were written in the MATLAB programming language using Matlab 5.3 (MathWorks Inc, Natick, MA, USA). Mean centering was applied to the data only for GPLS.

Results and Discussion

The data set used in this study was selected to demonstrate the applicability of 4 new genetic algorithm based multivariate calibration methods to simultaneous determination of multicomponent mixtures based on NIR spectroscopy. Methylene chloride and ethyl acetate were selected based on the similarities in their spectra whereas methanol selection was based on its dominating O–H overtone bands, which mostly cover the other 2 components. Figure 1 shows the NIR spectra of pure components between 1000 and 2000 nm. As can be seen, methylene chloride and methanol show somewhat similar absorbance peaks around 1400 and 1700 nm whereas methanol gives broad absorbance bands around these regions, resulting in a complete overlap and domination by the methanol. In addition to this complete methanol domination, the maximum absorbance in these regions is greater than 5, which results in non-linearity in the spectra. Using a smaller pathlength sample holder would have solved this non-linearity problem, but we planned to have some regions with a non-linearity problem in the spectra so that it would be possible to evaluate the behavior of each genetic method in these regions.

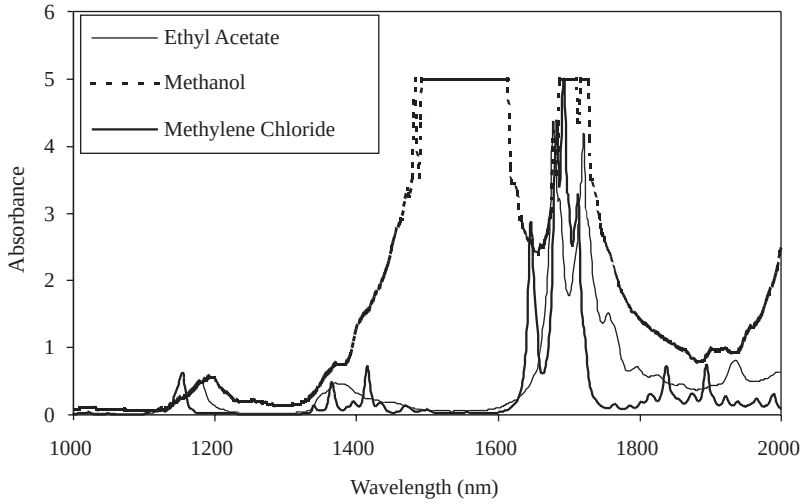

Figure 1. Near infrared spectra of pure components of the ternary system between 1000 and 2000 nm.

Figure 2 illustrates the NIR spectra of 3 ternary mixtures between 1000 and 2000 nm. Here the maximum non-linearity is observed from 1650 to 1750 nm. It is also interesting to note that the region between 1400 and 1600 nm is mostly determined by the methanol content of the sample. Overall, some minute differences exist on the spectra due to the different percentage compositions of the mixtures, and throughout the multivariate calibration process, it is expected that these differences will reveal the information necessary to build successful calibration models otherwise almost impossible with univariate calibration methods.

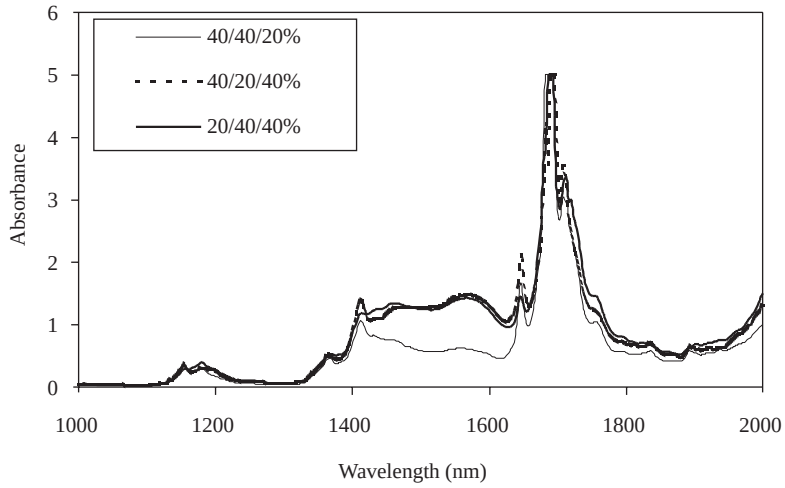

Figure 2. Near infrared spectra of ternary mixtures between 1000 and 2000 nm. The mixing order of the components in the legend is given as methylene chloride / ethyl acetate / methanol.

Standard error of calibration (SEC) and standard error of prediction (SEP) results for the calibration and the validation sets obtained with the GR method for the components of the ternary system are shown in Figure 3a for the model with 21 samples and Figure 3b for that with 42 samples. The SEC and SEP values for methylene chloride and methanol in both designs ranged from 0.62 to 1.43% whereas the results for ethyl acetate were between 1.90 and 2.57%. Figure 4 illustrates the SEC and SEP results obtained with the GPLS method for all the components of the mixtures in the first and second designs. Here, the SEC and SEP values for the 3 components ranged between 0.59 and 1.96%, where methylene chloride gave the smallest errors and overall, results of GPLS were similar to those of GR.

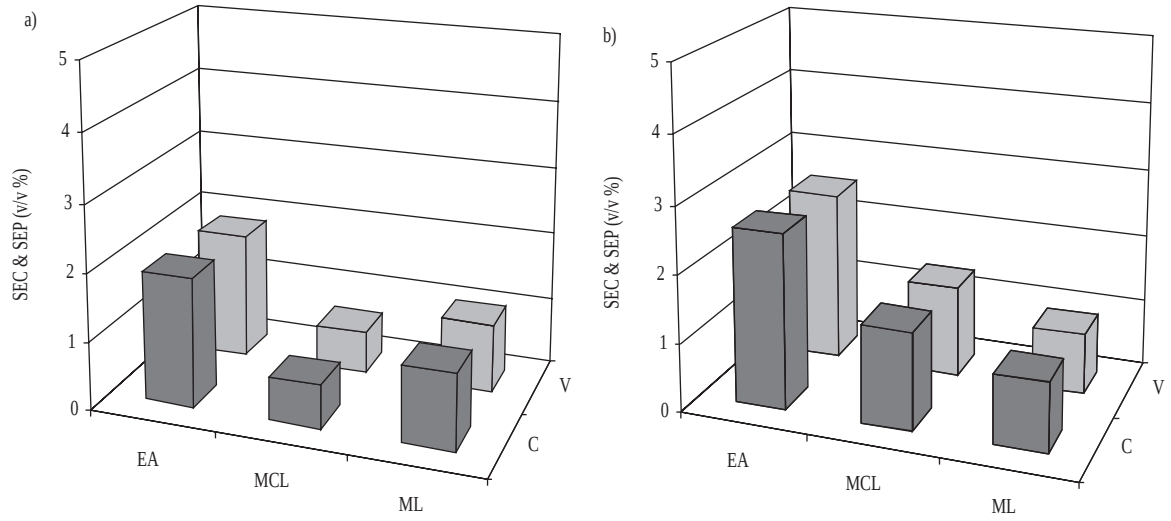

Figure 3. Standard error of calibration (SEC) and standard error of prediction (SEP) results for the calibration and validation sets obtained with the GR method for the components of the ternary system. a) Calibration set with 21 samples, b) Calibration set with 42 samples. (EA; ethyl acetate, MCL; methylene chloride, ML; methanol, C; calibration, V; validation).

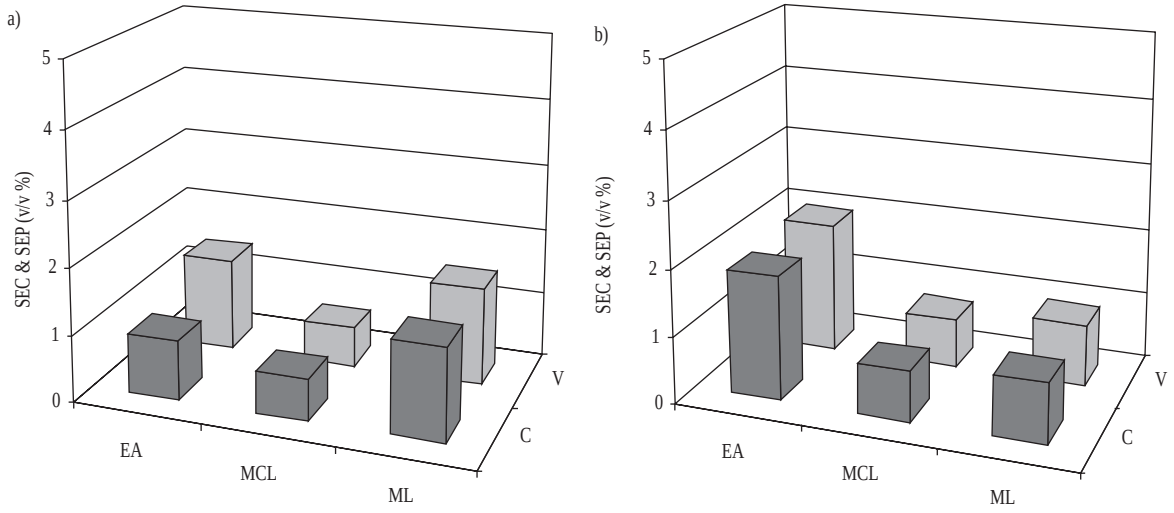

Figure 4. Standard error of calibration (SEC) and standard error of prediction (SEP) results for the calibration and validation sets obtained with the GPLS method for the components of the ternary system. a) Calibration set with 21 samples, b) Calibration set with 42 samples. (EA; ethyl acetate, MCL; methylene chloride, ML; methanol, C; calibration, V; validation).

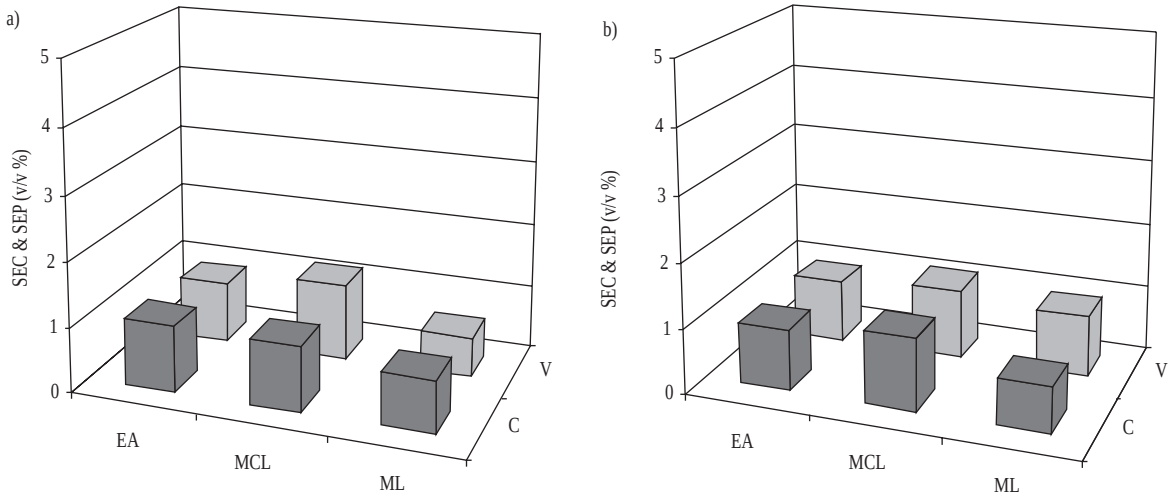

Figure 5. Standard error of calibration (SEC) and standard error of prediction (SEP) results for the calibration and validation sets obtained with the GCLS method for the components of the ternary system. a) Calibration set with 21 samples b) Calibration set with 42 samples. (EA; ethyl acetate, MCL; methylene chloride, ML; methanol, C; calibration, V; validation).

The SEC and SEP results for the calibration and the validation sets obtained with the GCLS method are shown in Figure 5a for the model with 21 samples and Figure 5b for that with 42 samples. The GCLS method produced similar SEC and SEP values for methanol and methylene chloride and relatively lower results for methylene chloride compared to the GR and GPLS results for the same component. This can be explained by the fact that the mixtures are well defined for the model building step of CLS and no interfering species are present. The results of the last genetic multivariate calibration method are presented

in Figure 6 where the SEC and SEP results of GILS ranged from 0.22 to 0.66% for all the components. Compared to the above 3 methods, some reduction in SEC and SEP values can be seen for GILS, which may be explained by the fact that ILS can be a powerful multivariate calibration method when accompanied by a proper wavelength selection procedure. Table 2 illustrates the SEC and SEP results for all methods and components of the mixtures along with a column illustrating the correlation coefficients (R^2) of actual vs. NIR predicted component concentrations plots. The R^2 values ranged from 0.9928 to 1.000 as shown in Figure 7, which gives the actual vs. the GILS predicted component concentrations for the 3 components in 21 calibration samples.

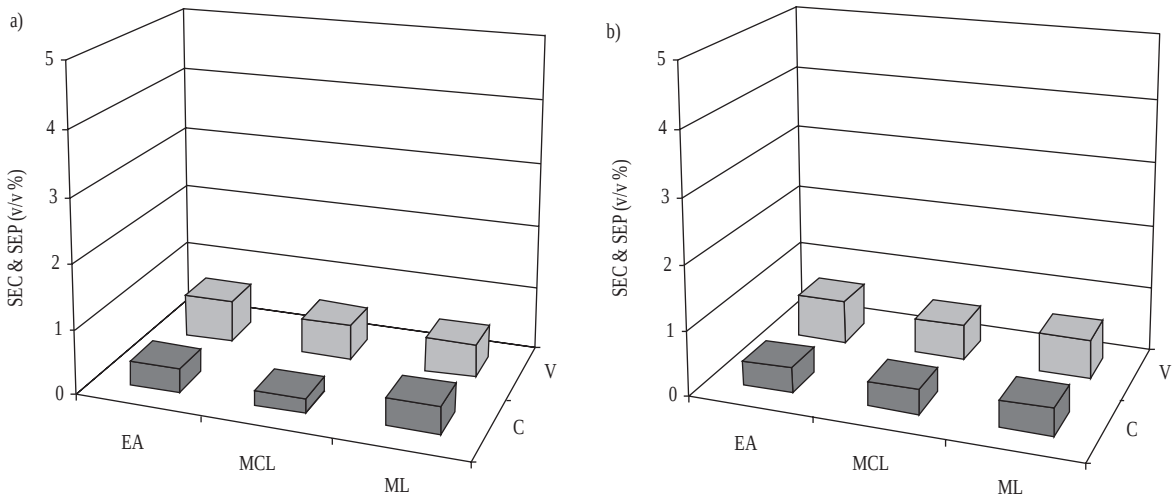

Figure 6. Standard error of calibration (SEC) and standard error of prediction (SEP) results for the calibration and validation sets obtained with the GILS method for the components of the ternary system. a) Calibration set with 21 samples, b) Calibration set with 42 samples. (EA; ethyl acetate, MCL; methylene chloride, ML; methanol, C; calibration, V; validation).

Table 2. The SEC, SEP and R^2 results for all the components and methods.

Name of Method	Components	Models with 21 Calibration Samples			Models with 42 Calibration Samples		
		SEC (v/v %)	SEP (v/v %)	R^2	SEC (v/v %)	SEP (v/v %)	R^2
GR	EA	1.9	1.85	0.9963	2.57	2.49	0.9929
	MLC	0.64	0.62	0.9996	1.43	1.34	0.9978
	ML	1.11	1.01	0.9987	1.02	0.90	0.9989
GPLS	EA	0.89	1.39	0.9992	1.86	1.96	0.9963
	MLC	0.62	0.59	0.9996	0.75	0.73	0.9994
	ML	1.39	1.47	0.9980	0.89	0.92	0.9991
GCLS	EA	1.02	0.94	0.9989	0.94	0.94	0.9990
	MLC	1.01	1.18	0.9990	1.10	1.07	0.9987
	ML	0.80	0.62	0.9993	0.69	0.94	0.9995
GILS	EA	0.36	0.66	0.9999	0.39	0.66	0.9998
	MLC	0.22	0.54	1.0000	0.38	0.56	0.9998
	ML	0.43	0.52	0.9998	0.41	0.60	0.9998

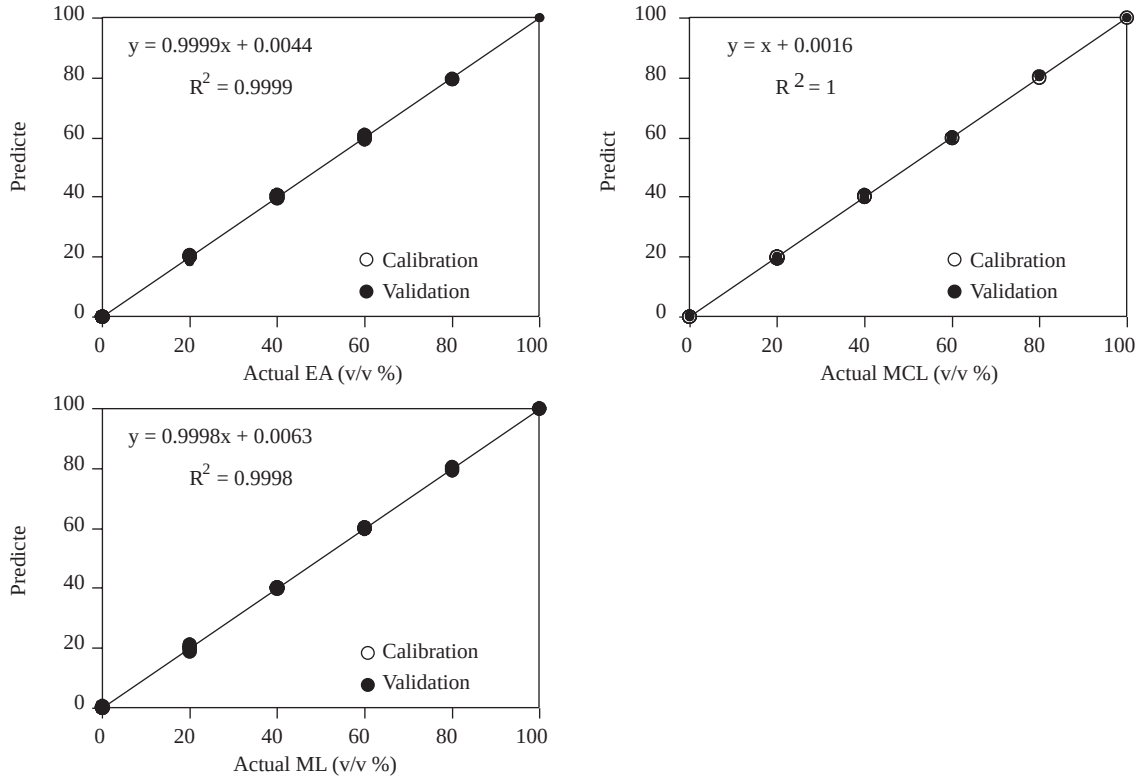

Figure 7. Plots of actual vs. predicted component concentrations for the calibration and validation sets obtained with the GILS method. (EA; ethyl acetate, MCL; methylene chloride, ML; methanol, C; calibration, V; validation).

The GPLS is a factor based method and GR is a hybrid method between the univariate and multivariate calibration approaches. In the case of GPLS, a small subset of wavelengths is selected based on the success of the first PLS factor which is reasonable when one considers that most of the variability in the data is contained in the first principle component of the data. The GCLS method is an extension of the normal CLS method which assumes that the absorbance at each wavelength is directly proportional to the component concentrations that make up the mixtures. The advantage with GCLS is that the genetic algorithm used in this method eliminates the wavelengths that are not proportional to the concentrations of the components.

On the other hand, the GR method works by taking the linear combinations of some genetically selected wavelengths using simple mathematical operators (+, –, *, /). In this respect, these methods are quite different yet the 3 methods were able to generate similar results. Figures 8–10 illustrate the distribution of the genetic algorithm selected wavelengths for all the components. As can be seen from these Figures, the best genes for GR, GPLS, and GCLS contain wavelengths from very similar regions, mostly concentrated in the 1200 nm region for methylene chloride and ethyl acetate. Genes for methanol have wavelengths from a broader region, which is expected since methanol is the component that dominates the spectra of the mixtures. The most interesting feature of the genes is that genetically selected wavelengths are not present in the final best genes where absorbance exceeds 2, even though these regions also had the same equal chance to be selected by the genetic algorithm in the initial step. One possible explanation of this is that these wavelengths were most probably selected at the initial random selection step but could not survive when the genetic algorithm started breeding and at the single point crossover steps.

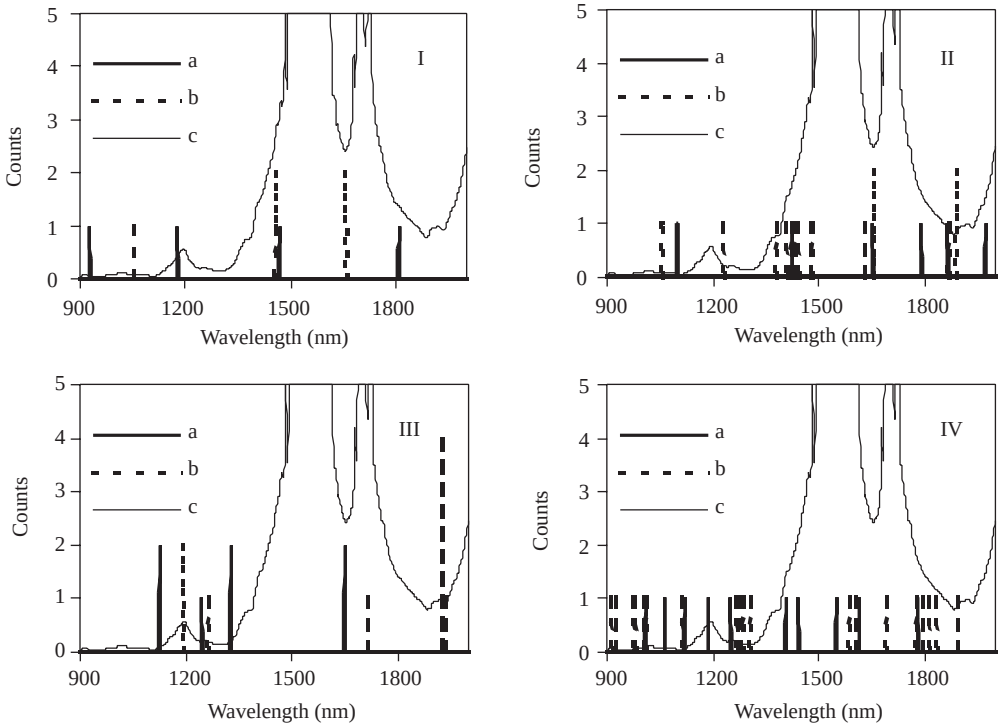

Figure 8. Distribution of the genetic algorithm selected wavelengths for methanol: I) GR, II) GPLS, III) GCLS, and IV) GILS (a; models with 21 calibration samples, b; models with 42 calibration samples, c, pure component spectrum of methanol).

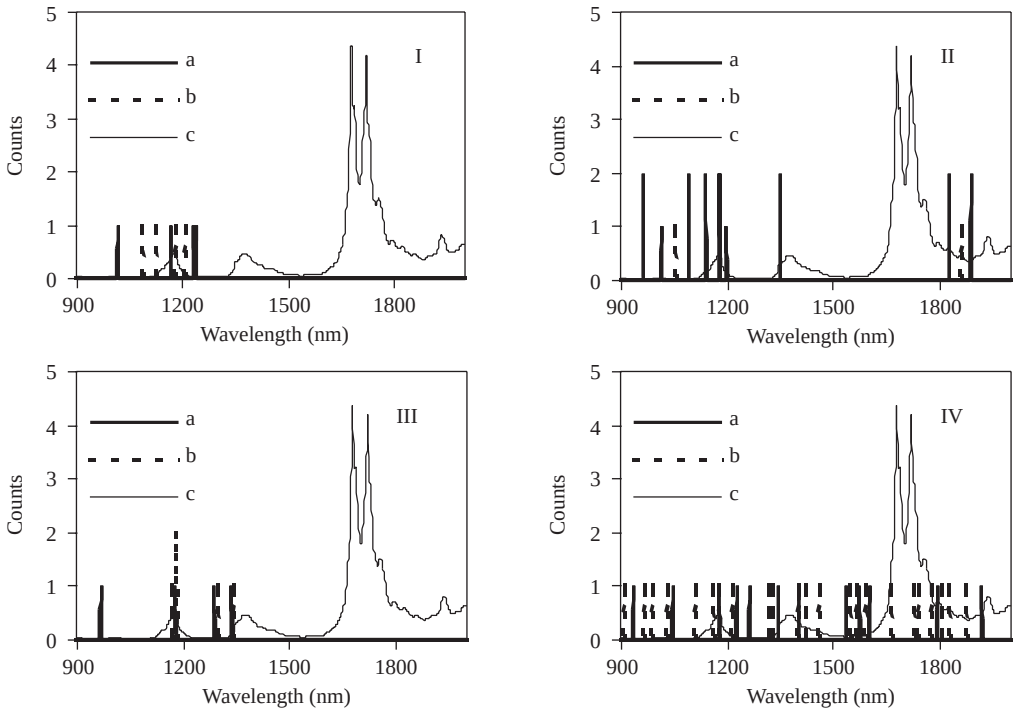

Figure 9. Distribution of the genetic algorithm selected wavelengths for methyl acetate I) GR, II) GPLS, III) GCLS, and IV) GILS (a; models with 21 calibration samples, b; models with 42 calibration samples, c; pure component spectrum of methanol).

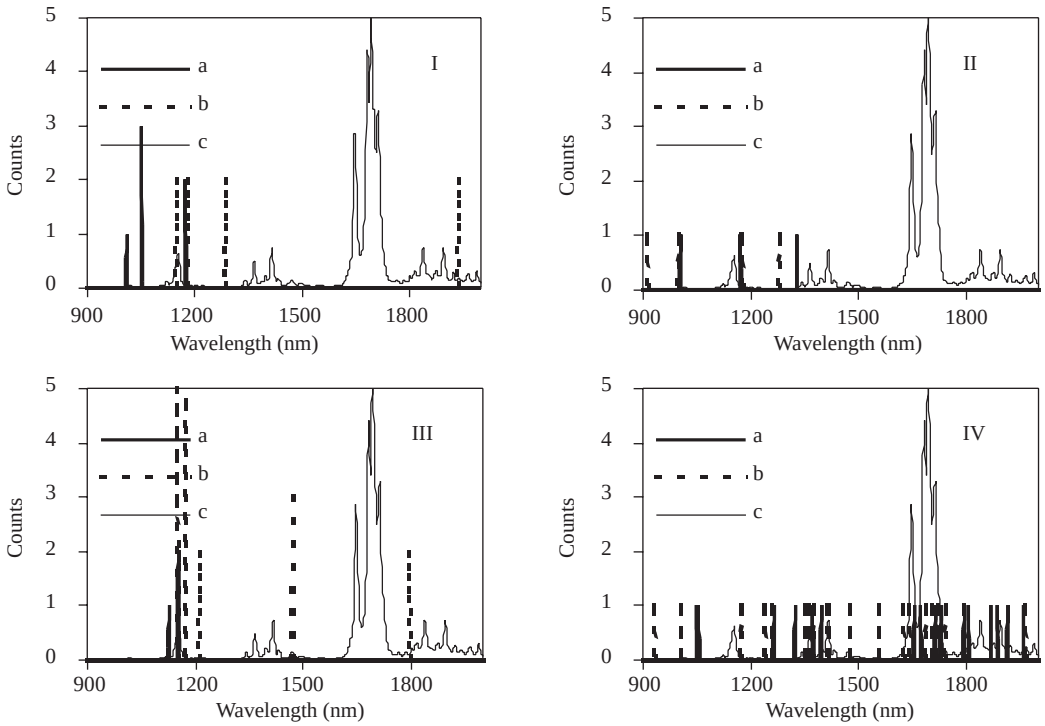

Figure 10. Distribution of the genetic algorithm selected wavelengths for methylene chloride I) GR, II) GPLS, III) GCLS, and IV) GILS (a; models with 21 calibration sample, b; models with 42 calibration samples, c; pure component spectrum of methanol).

The most interesting results obtained in this study were the results of GILS in which the best genes for the 3 components were quite different from those obtained from other 3 methods, yet the SEC and SEP values are better than those of GR, GPLS and GCLS even though some of the wavelengths in the genes are selected from high absorbance regions. Figure 11 shows the distribution of the genetic algorithm selected wavelengths for all the components and methods, along with a spectrum of a ternary mixture having 40% methylene chloride 40% ethyl acetate and 20% methanol by volume. As can be seen from the figure, wavelengths around 1200 and 1400 nm were selected several times which indicates the highest proportionality between concentration and absorbance in this regions.

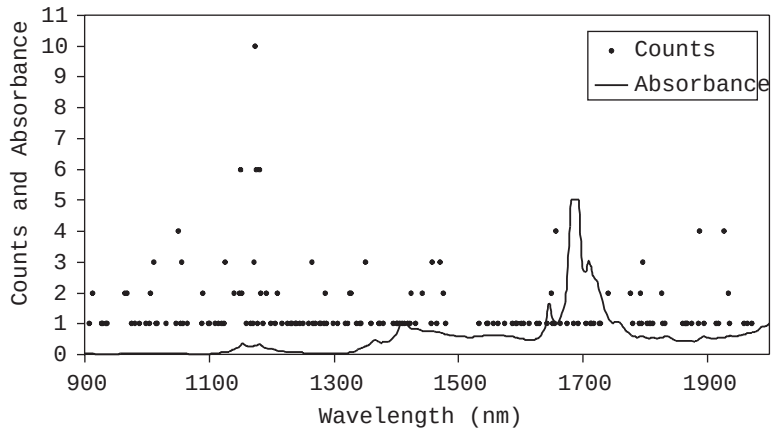

Figure 11. Distribution of the genetic algorithm selected wavelengths for all the components and methods along with a spectrum of a ternary mixture having 40% methylene chloride, 40% ethylacetate and 20% methanol by volume.

The genetically selected wavelengths for the best modes generated with the 4 methods are illustrated in Tables 3 and 4 for the calibration set with 21 and 42 sample sets, respectively. As can be seen from the Table, the GR method needed a minimum of 4 and a maximum of 8 specific wavelengths to generate a successful calibration model. Even though these are very small genes with only 2 to 4 base pairs, we observed that, most of the time, the successful models were those with a small number of base pairs (usually between 2 and 5). The most likely reason for the relatively small gene size is the over fitting problem as in all multivariate calibration methods. When the gene becomes larger and larger each additional base pair is actually fitting the small variations in the calibration set but these variations may not be present in the prediction set.

Table 3. Genetically selected wavelengths for the best models generated with the 4 methods in the case of 21 calibration samples. (EA; Ethyl Acetate, MCL; Methylene Chloride, ML; Methanol, G; Gene).

Method	Component	Selected wavelengths (nm)
GR	EA	G = (1238 * 1016) + (1232 – 1166)
	MCL	G = (1050 + 1172) + (1050 + 1172) + (1010 – 1050)
	ML	G = (928 – 1178) + (1466 + 1808)
GPLS	EA (4)#	G : 1826, 964, 1090, 1140, 1350, 1888, 1176, 1198, 1014, 1180, 1826, 964, 1180, 1090, 1140, 1350, 1888, 1176
	MCL (3)#	G : 11721, 1328, 1004
	ML (4)#	G : 1972, 1098, 1424, 1652, 1864, 1100, 1790
GCLS	EA	G : 1336, 1176, 1288, 968
	MCL	G : 1150, 1152, 1152, 1126
	ML	G : 1326, 1650, 1126, 1126, 1246, 1326, 1650
GILS	EA	G : 936, 1226, 1344, 1402, 1602, 1176, 1534, 1792, 1574, 1046, 1264, 1918
	MCL	G : 1264, 1958, 1864, 1372, 1674, 1654, 1056, 1914, 1792, 1886, 1396, 1050, 1728, 1322, 1712, 1806
	ML	G : 1442, 1780, 1406, 1614, 1250, 1186, 1548, 1064, 1120

The optimum number of PLS factors for this particular gene.

Table 4. Genetically selected wavelengths for the best models generated with the 4 methods in the case of 42 calibration samples. (EA; Ethyl Acetate, MCL; Methylene Chloride, ML; Methanol, G; Gene).

Method	Component	Selected wavelengths (nm)
GR	EA	G = (1206 – 1180) + (1124 * 1088)
	MCL	G = (1934 * 1286) + (1180 – 1148) + (1934 * 1286) + (1180 – 1148)
	ML	G = (1456 + 1054) + (1656 + 1458) + (1656 + 1458)
GPLS	EA (3)#	G : 1056, 1180, 1860
	MCL (4)#	G : 1176, 912, 1280, 928
	ML (3)#	G : 1656, 1888, 1432, 1476, 1380, 1056, 1058, 1230, 1630, 1422, 1656, 1480, 1866, 1410, 1442, 1888
GCLS	EA	G : 1182, 1342, 1170, 1298, 1182
	MCL	G : 1470, 1150, 1174, 1174, 1174, 1210, 1470, 1150, 1796, 1150, 1174, 1174, 1174, 1174, 1210, 1470, 1150, 1796, 1150, 1174, 1174
	ML	G : 1936, 1192, 1926, 1264, 1192, 1926, 1926, 1926, 1714
GILS	EA	G : 1160, 968, 1740, 1594, 908, 1216, 1328, 1564, 1804, 1828, 1424, 1726, 1458, 1030, 1110, 1874, 1592, 1664, 1400, 1318, 988, 1176,
	MCL	G : 1718, 934, 1686, 1704, 1626, 1964, 1360, 1556, 1476, 1740, 1414, 1374, 1350, 1260, 1006, 1642, 1240, 1174, 1796, 1418
	ML	G : 1794, 1692, 1832, 912, 1270, 1776, 976, 1306, 1284, 1010, 980, 926, 1006, 1604, 1894, 1114, 1586, 1278, 1266, 1812

The optimum number of PLS factors for this particular gene.

When comparing genetic algorithm selected wavelengths with the spectra of pure components given in Figure 1, some of the wavelengths are from the nonlinear regions of the spectra. The most likely reason for this is that all the algorithms are completely based on random processes. It is also interesting to note that calibration models with 42 samples generated relatively higher SEC and SEP values. The reason for this trend could be the that all the sample sets with 21 spectra used in this study were prepared at different times, so there could be day to day variations between the spectra of each set.

The numbers of selected wavelengths for GPLS and GCLS are also surprisingly small, ranging between 3 and 21. In addition, the optimum number of PLS factors varied between 3 and 4 even for the larger genes. There were many other genes that had greater or smaller wavelengths than these values, but these were the ones that produced the best SEC and SEP results. Finally, the number of wavelengths in the best genes of the GILS method is larger than those of the 3 methods, ranging between 9 and 22.

Conclusions

This study shows that wavelength selection based on a genetic algorithm can improve the accuracy of hard modeling multivariate calibration techniques (ILS and CLS) for NIR spectra. In fact, the GILS method was able to yield the lowest SEP values for this particular NIR data set. It is also interesting to note that the GR method was able to give standard calibration and prediction errors in the range of the other 3 methods even though GR generates models based on the simple least squares procedure. In terms of model building efficiency, it seems that there is not great difference among these 4 methods. However, GILS might be preferred over GR and PLS since GR and PLS are more computationally intensive methods.

References

1. D.A. Burns and E.W. Ciurczak, "**Handbook of Near-Infrared Analysis**", Marcel Dekker, New York, 1992.
2. W.F. McClure, Anal. Chem. 66, 43A–53A, (1994).
3. F.A. DeThomas, J.W. Hall, and S.L. Monfre, **Talanta 41**, 425–31 (1994).
4. R. Raghavachari, "**Near Infrared Applications in Biotechnology**", Marcel Dekker, New York, 2001.
5. S.A. Arnold, J. Crowley, S. Vaidyanathan, L. Matheson, P. Mohan, J. W. Hal, L. M. Harvey, and B. McNeil, **Enzyme and Microbial Technology 27**, 691–7 (2000).
6. E.W. Ciurczak and J. K. Drennen, "**Pharmaceutical and Medical Applications of Near-Infrared Spectroscopy**", Marcel Dekker, New York, 2002.
7. U. Wählby, and J. Skjöldebrand, **J. Food Eng. 47**, 303–12 (2001).
8. R.P. Paradkar,and R.R. Williams, **Appl. Spectrosc. 51**, 92–100 (1997).
9. D. Özdemir, R.M. Mosley, and R.R. Williams, **Appl. Spectrosc. 52**, 599–603 (1998).
10. D. Özdemir, R.M. Mosley, and R.R. Williams, **Appl. Spectrosc. 52**, 1203–9 (1998).
11. R.M. Mosley, and R.R. Williams, **Appl. Spectrosc. 52**, 1197–202 (1998).
12. D. Özdemir, and R. R. Williams, **Appl. Spectrosc. 53**, 210–7 (1999).
13. D. Lawerence, "**Handbook of Genetic Algorithms**", Van Nostrand Reinhold, New York, 1991.

14. C.B. Lucasius, and G. Kateman, **Chem. Intell, Lab. Syst. 19**, 1–33 (1993).
15. U. Hörchner, J.H. Kalivas, **Anal. Chim. Acta 311**, 1–13 (1995).
16. D.M. Haaland, and EV. Thomas, **Anal Chem. 60**, 1193–202 (1988).
17. P. Geladi, and B.R. Kowalski, **Anal. Chim. Acta 185**, 1–17 (1986).
18. P.D. Wentzell, D.T. Andrews, and B.R. Kowalski, **Anal. Chem. 69**, 2299–311 (1997).
19. K. Esbensen, P. Geladi, and S. Wold, **Chem. Intell. Lab. Syst. 2**, 37–52 (1987).