Determination of Octane Number of Gasoline Using Near Infrared Spectroscopy and Genetic Multivariate Calibration Methods

Durmuş Özdemir

Published online: 14 Feb 2007.

Submit your article to this journal ⬀

Article views: 86

View related articles ⬀

Citing articles: 9 View citing articles ⬀

Taylor & Francis
Taylor & Francis Group

Determination of Octane Number of Gasoline Using Near Infrared Spectroscopy and Genetic Multivariate Calibration Methods

Durmuş Özdemir

Department of Chemistry, Faculty of Science, İzmir Institute of Technology,
Gülbahçe, Urla, İzmir, Turkey

Abstract: The feasibility of rating the octane number of gasoline using near infrared (NIR) spectroscopy and three different genetic algorithm-based multivariate calibration methods was demonstrated. The three genetic multivariate calibration methods are genetic regression (GR), genetic classical least squares (GCLS), and genetic inverse least squares (GILS). The sample data set was obtained from the ftp address (*ftp://ftp.clarkson.edu/pub/hopkepk/Chemdata/*) with the permission of Professor. J. H. Kalivas. This data set contains the NIR spectra of 60 gasoline samples collected using diffuse reflectance as $\log (1/R)$ with known octane numbers and covers the range from 900 to 1700 nm in 2 nm intervals. Of these 60 spectra, 20 were used as the calibration set, 20 were used as the prediction set, and 20 were reserved for the validation purposes. Several calibration models were built with the three genetic algorithm-based methods, and the results were compared with the partial least squares (PLS) prediction errors reported in the literature. Overall, the standard error of calibration (SEC), standard error of prediction (SEP), and standard error of validation (SEV) values were in the range of 0.15–0.32 (in the units of motor octane number) for the GR and GILS, which are comparable with the literature. However, GCLS produced relatively large results (0.36 for SEC, 0.39 for SEP and 0.52 for SEV) when compared with the other two methods.

INTRODUCTION

Gasoline is a highly complex mixture of various hydrocarbons and oxygenates, such as methyl t-butyl ether (MTBE). Therefore, the resulting spectrum of gasoline is the summation and highly overlapped spectral features

of each component in gasoline. The spectral features of gasoline are dominated by overtone and combination bands of CH, CH_2, and CH_3, groups, and the most useful spectral information is located in the 1100–1650 nm and 1800–2100 nm spectral ranges (Hoeil, Hyeseon, and Chi Hyuck, 2001). There have been a number of reports in the past years about the prediction of octane number of gasoline using spectroscopic techniques (Kelly and Callis, 1990; Fodor, Kohl, and Mason, 1996; Litani-Barzilai et al., 1997; Kalivas, 1997; Brenchley, Hörchner, and Kalivas, 1997). However, the conventional method used to determine the octane number in the petroleum industry is the knocking engine, which requires long analysis times and is not suitable for online monitoring.

Near infrared (NIR) spectroscopy (McClure, 1994) has become a popular method for simultaneous chemical analysis and is being studied extensively in a number of different fields, such as process monitoring (DeThomas, Hall, and Monfre, 1994), biotechnology (Arnold et al., 2000), and pharmaceutical and food industry (Wählby and Skjöldebrand, 2001), because of the potential for online, 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 vibrational bands. Although all the fundamental vibrational 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 led to the development of several multivariate calibration methods (DeThomas et al., 1994; Arnold et al., 2000; Wählby and Skjöldebrand, 2001) for the analysis of complex chemical mixtures.

Genetic regression (GR) (Paradkar and Williams, 1997; Özdemir, Mosley, and Williams, 1998a, 1998b; Mosley and Williams, 1998; Özdemir and Williams, 1999) is a calibration technique that optimizes linear regression models using a genetic algorithm (GA) and has been applied to a number of multi-instrument calibration and wavelength selection problems. GAs (Lucasius and Kateman, 1993; Hörchner and Kalivas, 1995) are nonlocal search and optimization methods that are based upon the principles of natural selection. For given full spectrum 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 original CLS and ILS (Haaland and Thomas, 1988; Geladi and Kowalski, 1986) methods in which a small set of wavelengths is selected from a full spectral data matrix and evolved to an optimum solution using a GA. The CLS, ILS, and PLS methods were well described by Haaland and Thomas (1988); Geladi and Kowalski (1986); Wentzell, Andrews, and Kowalski (1997); and Esbensen, Geladi, and Wold (1987).

In this work, three different genetic multivariate calibration methods, GR, GCLS, and GILS, were tested with the aim of establishing calibration models that have a high predictive capacity for the determination of the octane number of gasoline using the NIR spectroscopic technique.

THEORY

Genetic Regression

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 simple and consists of five 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.

GR is an implementation of a GA for selecting wavelengths and mathematical operators to build linear calibration models. GR is a hybrid calibration between univariate and multivariate calibration techniques in which it 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 GAs, but it 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 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 for the 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 two in order to allow single-point crossover, whereas the higher limit was set to eliminate overfitting problems and reduce 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).

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. Also, 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 gene mates with the second gene and the third one with the fourth one and so on, as illustrated in the following example:

Parents:

$$S_1 = (A_{347} * A_{251})\# + (A_{379} + A_{218}) \tag{1}$$

$$S_2 = (A_{225} * A_{478})\# + (A_{343}/A_{250}) + (A_{451} - A_{358}) + (A_{231} - A_{458}) \tag{2}$$

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

Offspring:

$$S_3 = (A_{347} * A_{251}) + (A_{343}/A_{250}) + (A_{451} - A_{358}) + (A_{231} - A_{458}) \tag{3}$$

$$S_4 = (A_{225} * A_{478}) + (A_{379} + A_{218} \tag{4}$$

Here, the first part of the S_1 is combined with the second part of the S_2 to give the S_3, and likewise, the second part of the S_1 combined with the first part of the S_2 to give S_4. This process is called the single-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 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 are reached.

Mutation that introduces random deviations into the population was also introduced into the GR during the mating step at a rate of 1%, as is typical in GAs. This is usually done by replacing one of the base pairs in an existing gene with a randomly generated new base pair. Mutation allows the GR to explore the search space and incorporate new material into the genetic population. It helps keep the search moving and can eject GR from a local minimum on the response surface. However, it is important not to set the mutation rate too high, because it may keep the GA from being able to exploit the existing population.

In the end, the gene with the lowest SEC (highest fitness) is selected for the model building, which is done by simple least squares. This model is used to predict the concentrations of the 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). Because the random processes are heavily involved in the GR as in all the GAs, the program has been set to run several times for each component in a given multicomponent mixture analysis. 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 with 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 full spectral information and reduces it to a single score upon which simple calibration models are built. First, 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, because 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. Note that data pretreatment is not necessary before calibration, which saves extra time in data processing.

Genetic Classical Least Squares

The 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 they were in the classical univariate method. In matrix notation,

the CLS model for m calibration samples containing l chemical components with spectra that contain n wavelengths is described as follows:

$$A = CK + E_A \tag{5}$$

where A is the $m \times n$ matrix of the calibration spectra; C is the $m \times l$ matrix of the component concentrations; K is the $l \times n$ matrix of absorptivity-pathlength constants; and E_A is the $m \times n$ matrix of the spectral errors or residuals not fit by the model. Here, the K matrix represents the first-order estimates of the pure component spectra at unit concentration and unit path length. The method of least squares can be used to estimate the K matrix. The least-squares estimate of the K is defined as follows:

$$\hat{K} = (C'C)^{-1}C'A \tag{6}$$

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

$$\hat{c} = (\hat{K}\hat{K}')^{-1}\hat{K}a \tag{7}$$

where a is the spectrum of the unknown sample, and $\hat{c}$ is the vector of the predicted component concentrations.

GCLS is a modified version of the original CLS method in which a small set of wavelengths is selected from full spectral data using a genetic algorithm. The algorithm used to select the optimum number of wavelengths in GCLS is similar to the GR algorithm but differs in the way it encodes the gene. In GCLS, the term "gene" describes a vector with elements that are randomly selected wavelengths. 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 a full 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 one with the fourth one and so on as described above with one difference. Because 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 one 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 are 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 is replaced with 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 the component being analyzed in the validation (test) sets as described in GR.

Genetic Inverse Least Squares

The major drawback of the 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 that 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

$$C = AP + E_C \tag{8}$$

where C and A are the same as in CLS, P is the $n \times l$ matrix of the unknown calibration coefficients relating l component concentrations to the spectral intensities, and E_C is the $m \times l$ 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 biggest advantage of ILS is that Eq. (8) can be reduced for the analysis of a single component at a time, as analysis based on an ILS model is invariant with respect to the number of chemical components included in the analysis. The reduced model is given as follows:

$$c = Ap + e_c \tag{9}$$

where c is the $m \times 1$ vector of concentrations for the analyte that is being analyzed, p is the $n \times 1$ vector of calibration coefficients, and e_c is the $m \times 1$ vector of concentration residuals not fit by the model. During the calibration step, the least-squares estimate of p is

$$\hat{p} = (A'A)^{-1}A' \cdot c \tag{10}$$

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

$$\hat{c} = a' \cdot \hat{p} \tag{11}$$

where $\hat{c}$ is the scalar estimated concentration, and 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 identities of interfering species still need to be known to prepare a good calibration sample set.

The major disadvantage of ILS can be seen in Eq. (10) where the matrix, which must be inverted, has dimensions equal to the number of wavelengths in the spectrum, and this number cannot exceed the number of calibration samples. This is a big restriction, because the number of wavelengths in a spectrum will generally be more 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 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 A in the 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 in 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 smaller numbers 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. As can be seen, the genetic algorithm used in GR, GCLS, and GILS is not only used to select a set of wavelengths but also to put them in a competition through an evolving algorithm.

EXPERIMENTAL

Instrumentation

The NIR spectra of gasoline samples used in this study were obtained from the ftp server (*ftp://ftp.clarkson.edu/pub/hopkepk/Chemdata/*) with the permission of Professor J. Kalivas. The details of the data set can be found elsewhere (Kalivas, 1997; Brenchley et al., 1997). This data set consists of 60 gasoline samples with specified octane numbers. Samples were measured using diffuse reflectance as log $(1/R)$ from 900 to 1700 nm in 2 nm intervals, and no further data pretreatment was applied. Of these 60 spectra, 20 were used as the calibration set, 20 were used as the prediction set, and 20 were reserved for the validation purposes. The octane numbers of these samples are shown in Table 1 for calibration, prediction, and validation sets, respectively.

Table 1. Octane numbers of the 60 gasoline samples used in calibration, prediction, and validation sets

Sample number	Octane number		
	Calibration set	Prediction set	Validation set
1	85.3	85.25	88.45
2	83.4	88.9	85.5
3	87.9	88.3	88.45
4	88.7	88.25	88.75
5	88	87.3	88.6
6	85.5	88.7	87.05
7	88.75	88.65	87.25
8	87.15	85.4	86.85
9	88.65	87	86
10	86.3	86.6	86.1
11	84.4	86.5	84.6
12	84.5	84.7	88.1
13	88.2	88.4	85.25
14	88.2	88.4	88.55
15	88	88.35	88.5
16	88.85	85.3	88.25
17	87.6	88.7	88.45
18	84.7	88.1	88.35
19	89.6	85.1	85.1
20	87.1	87.2	86.6

Software

The new GA-based multivariate calibration methods (GR, GCLS, and GILS) were written in MATLAB programming language using MATLAB 5.3 (Math-Works Inc., Natick, MA).

RESULTS AND DISCUSSION

The data set used in this study was selected to demonstrate the applicability of three new GA-based multivariate calibration methods to the octane number determination of gasoline based on NIR spectroscopy. Figure 1 shows the NIR spectra of three gasoline samples with specified octane numbers (84.7, 87.2, and 89.6). Because of structural similarities, the spectral features of these three samples are very much alike, and only minute differences exist in some parts of the whole spectrum. To illustrate these minute differences, a small portion of the spectra are also shown on the same figure, which
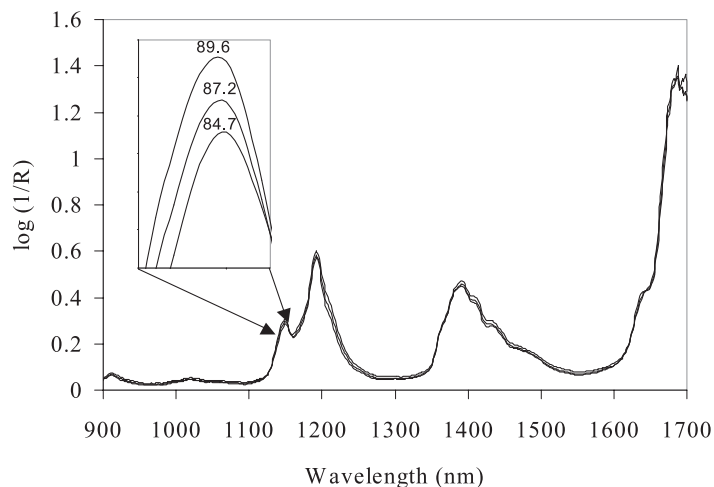
Figure 1. Near infrared spectra of three different gasoline samples collected in diffuse reflectance mode. The octane numbers of each are shown on top of each peak in the small figure. (Note that the peak around 1150 nm is enlarged in order to show the small absorbance difference between the three samples.)

is an enlarged view of the region between 1140 and 1155 nm. As can be seen from this magnified view of the spectra, there are small differences for these samples not only in the wavelength scale but also in the absorbance scale. 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.

The SEC, SEP, and SEV results obtained with the three GA-based methods (GCLS, GILS, and GR) are plotted in Figure 2 as a bar graph along with the corresponding values on each bar. All the GA-based methods except GCLS generated similar results with the earlier reported PLS results (Kalivas, 1997). Even though the results obtained with GR and GILS were comparable, it was observed that the SEC, SEP, and SEV values produced by GR were in better agreement compared to the results of GILS. On the other hand, the results obtained with GCLS were about twice those obtained with GR and GILS. One possible reason that GCLS did not work well with this octane data is that Beer's law applies to for the concentration of analyte, whereas in this study, it is the octane number of gasoline samples. Also, the CLS method requires that in order to produce successful calibration models, composition of the sample must be well defined, which is not the case here.

The genetically selected wavelengths for the best modes generated with the three methods are illustrated in Table 2. As can be seen from the table, the GR method needed only eight specific wavelengths to generate a successful calibration model. Even though this is a very small gene with only four base
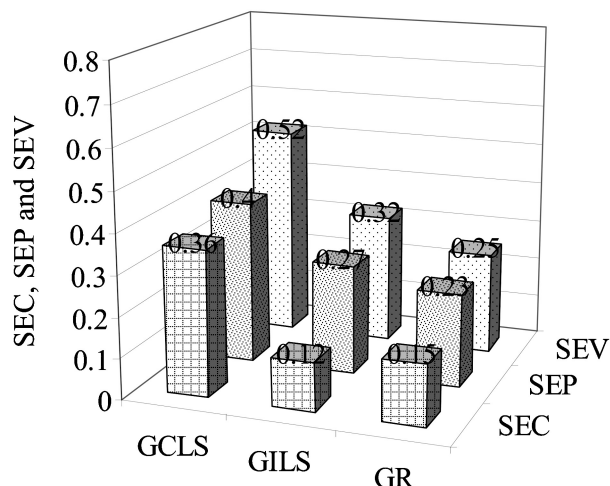
Figure 2. The standard error of calibration (SEC), standard error of prediction (SEP), and standard error of validation (SEV) results obtained from the three genetically modified calibration methods (GCLS, GILS, and GR).

pairs, we have observed that, for the most of the time, the successful models were the ones that have a small number of base pairs (usually between two and five) that are formed with the use of a subtraction operator. The most probable reason for the relatively small gene size is because of the over-fitting problem as in all multivariate calibration methods. When the gene gets larger, each additional base pair is actually fitting the small variations in the calibration set, but these variations may not present in the prediction set.

The number of selected wavelengths for GILS is 10 and for GCLS is 23. There were many other genes that had more or less wavelengths than these values, but these were the ones that produced the best SEC and SEP results. When one closely examines these wavelengths, notice that for the two methods, the GA selects similar wavelengths. Figure 3 shows the reference octane numbers versus the predicted octane numbers for the gasoline samples using

Table 2. Genetically selected wavelengths for the best modes generated with the GR, GILS, and GCLS methods

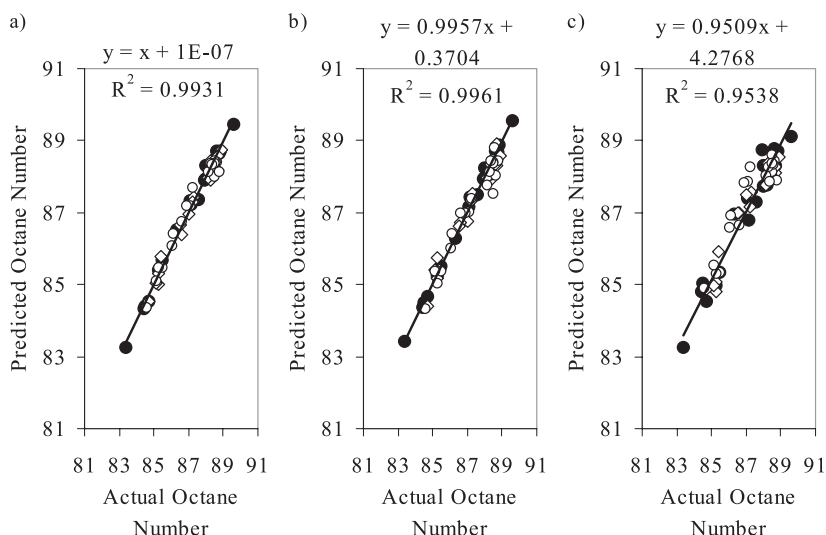| Name of the method	Selected wavelengths (nm)
GR	Gene = $(1230 - 1362) + (1636/1672) + (1464/1360) + (1208 + 1634)$
GILS	1148, 1180, 1522, 1002, 1118, 1434, 1462, 1582, 1214, 1374
GCLS	1250, 1092, 1556, 1582, 1328, 1336, 1582, 1008, 1146, 1136, 1378, 1106, 1096, 1250, 1092, 1556, 1336, 1582, 1278, 1106, 1086, 1336, 1582

Figure 3. Actual octane numbers versus the predicted octane numbers for the gasoline samples contained in calibration, prediction, and validation sets. (a) with GR, (b) with GILS, and (c) with GCLS (●: calibration set, ◇: prediction set, and ○: validation set).

three GA-based multivariate calibration methods. The regression coefficients for the GILS and GR were around 0.99 and for the GCLS was 0.95.

In order to determine whether the GA-selected wavelengths correspond to the particular information, the GR method was also set to run 50 times with 20 genes and 100 iterations. The overall distribution of the selected wavelengths along with a gasoline sample spectrum is shown in Figure 4 for the GR method. As can be seen from the figure, the GR method selected the wavelengths that correspond to certain regions of the spectrum that best relates the octane number to the spectral intensities. Notice that the algorithm starts with the whole spectrum information at the beginning of each run, and each wavelength has an equal chance of being selected. The explanation of this is in the evolutionary nature of the GA, where the wavelengths that are suited for the particular information survive in the long run of iterations and others do not. This gives an advantage to the GA-based methods, where only the information related to the particular component is used to construct the model, thereby reducing the noise in the overall information.

CONCLUSIONS

This study shows that wavelength selection based on a GA can improve the accuracy of hard modeling multivariate calibration techniques (ILS and CLS)
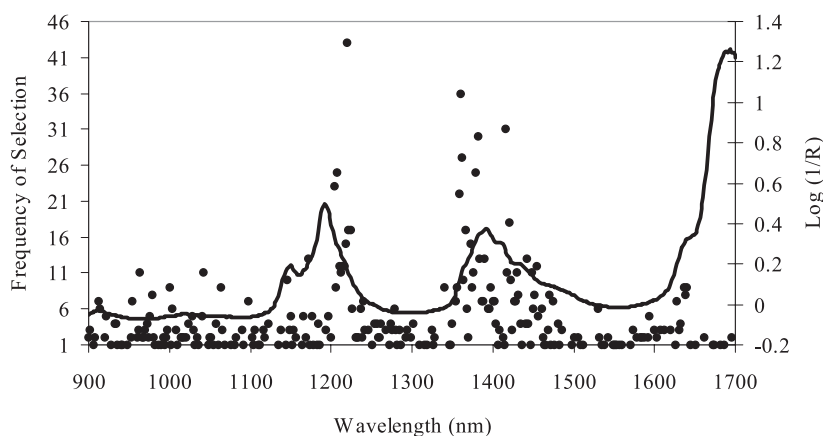
Figure 4. Distribution of the selected wavelengths by genetic algorithm in GR method for a total of 50 runs with 20 genes and 100 iterations along with a gasoline sample spectrum.

for NIR spectra. 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 three methods, even though the GR generates the models based on the simple least-squares procedure.

ACKNOWLEDGMENTS

REFERENCES

Arnold, S. A., Crowley, J., Vaidyanathan, S., Matheson, L., Mohan, P., Hall, J. W., Harvey, L. M., and McNeil, B. 2000. At-line monitoring of a submerged filamentous bacterial cultivation using near infrared spectroscopy. *Enzyme and Microbial Technology* 27(9):691–697.

Brenchley, J. M., Hörchner, U., and Kalivas, J. H. (1997). Wavelength selection characterization for NIR spectra. *Appl. Spectrosc.* 51(5):689–699.

DeThomas, F. A., Hall, J. W., and Monfre, S. L. (1994). Real-time monitoring of polyurethane production using near infrared spectroscopy. *Talanta* 41(3):425–431.

Esbensen, K., Geladi, P., and Wold, S. (1987). Principal component analysis. *Chem. Intell. Lab. Syst.* 2(1–3):37–52.

Fodor, G. E., Kohl, R. L., and Mason, R. L. (1996). Analysis of gasolines by FT-IR spectroscopy. *Anal. Chem.* 68(1):23–30.

Geladi, P., and Kowalski, B. R. (1986). Partial least squares regression: A tutorial. *Analytica. Chimica. Acta.* 185:1–17.

Haaland, D. M., and Thomas, E. V. (1988). Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of quantitative information. *Anal. Chem.* 60(11):1193–1202.

Hoeil, C., Hyeseon, L., and Chi-Hyuck, J. (2001). Determination of research octane number using NIR spectral data and ridge regression. *Bull. Korean Chem. Soc.* 22(1):37–42.

Hörchner, U., and Kalivas, J. H. (1995). Further investigation on a comparative study of simulated annealing and genetic algorithm for wavelength selection. *Analytica. Chimica. Acta.* 311(1):1–13.

Kalivas, J. H. (1997). Two data sets of near infrared spectra. *Chem. Intell. Lab. Syst.* 37(2):255–259.

Kelly, J. J., and Callis, J. B. (1990). Nondestructive analytical procedure for simultaneous estimation of the major classes of hydrocarbon constituents of finished gasolines. *Anal. Chem.* 62(14):1444–1551.

Litani-Barzilai, I., Sela, I., Bulatov, V., Zilberman, I., and Schechter, I. (1997). On-line remote prediction of gasoline properties by combined optical methods. *Analytica. Chimica. Acta.* 339(1–2):193–199.

Lucasius, C. B., and Kateman, G. (1993). Understanding and using genetic algorithms. Part 1. Concepts, properties and context. *Chem. Intell. Lab. Syst.* 19(1):1–33.

McClure, W. F. (1994). Near infrared spectroscopy—the giant is running. *Anal. Chem.* 66(1):43A–53A.

Mosley, R. M., and Williams, R. R. (1998). Determination of the accuracy and efficiency of genetic regression. *Appl. Spectrosc.* 52(9):1197–1202.

Özdemir, D., Mosley, R. M., and Williams, R. R. (1998a). Hybrid calibration models: An alternative to calibration transfer. *Appl. Spectrosc.* 52(4):599–603.

Özdemir, D., Mosley, R. M., and Williams, R. R. (1998b). Effect of wavelength drift on single and multi-instrument calibration using genetic regression. *Appl. Spectrosc.* 52(9):1203–1209.

Özdemir, D., and Williams, R. R. (1999). Multi-instrument calibration in UV-visible spectroscopy using genetic regression. *Appl. Spectrosc.* 53(2):210–217.

Paradkar, R. P., and Williams, R. R. (1997). Genetic regression as a calibration technique for solid phase extraction of dithizone-metal chelates. *Appl. Spectrosc.* 51(1):92–100.

Wählby, U., and Skjöldebrand, J. (2001). NIR measurements of moisture changes in foods. *Journal of Food Engineering* 47(4):303–312.

Wentzell, P. D., Andrews, D. T., and Kowalski, B. R. (1997). Maximum likelihood multivariate calibration. *Anal. Chem.* 69(13):2299–2311.