Refinery octane blend modelling using principal components regression of gas chromatography data

Norris R. Crawford and Walter W. Hellmuth
Texaco Research Center, PO Box 509, Beacon, NY 12508, USA
(Received 29 August 1989; revised 9 January 1990)

The use of capillary gas chromatography (g.c.) coupled with principal components regression (PCR) is shown to be an effective predictive tool for determining the octane values of gasolines blended from a wide variety of refinery streams. The linear combination of fuel stream g.c. data to give calculated blended fuel gas chromatograms and their associated octanes offers much promise for refinery blending optimization.

(Keywords: fuel; gas chromatography; gasoline)

Gas chromatographic data for gasolines and gasoline fractions have often been used to predict the physical and performance properties of these mixtures. The molecular structure and the volatility of gasoline components are both described precisely by the relative retention times observed in well controlled temperature programmed capillary g.c. runs. However, use of all the data captured by capillary g.c. to make octane predictions by a classical group additivity approach is extremely difficult, since complete characterization of up to 400 components is needed and it is nearly impossible to determine the contribution of each of the components to research and motor octane.

Assimilation of these components into 'group types' with proportionately different contributions to octane has been accomplished[1,2] with varying degrees of success[3]. In this work, the use of principal components regression is described as a means of using all of the data in a g.c. run to predict research (RON), motor (MON) and pump ((RON+MON)/2) octane numbers.

Clearly, a model constructed from multivariate capillary g.c. data using chemometric methodology would be advantageous and should be a better predictor of gasoline properties than similar approaches based on spectroscopic data alone. Both univariate (i.r.[4] and n.m.r.[5]) and multivariate (near i.r.[6]) regression techniques have been employed. Spectroscopic techniques address structural characteristics directly, but do not address volatility related parameters. Additionally, spectral shifts associated with matrix effects in both i.r. and n.m.r. data do not readily allow the calculation of mixture spectra from component spectra, for subsequent use in prediction of octanes. On the other hand, normalized g.c. data with good resolution can be combined linearly, to give a good representation of the mixture g.c. data, which can be used for prediction of octanes. Use of multivariate calibration techniques in model formation and prediction[7] allows all of the pertinent g.c. information to be used.

This work describes the building of a capillary g.c. based model to effectively cover the range of compositions encountered in a typical refinery consisting of eight different refinery streams. Use of this model within its design ranges is shown to accurately predict the octanes (RON, MON and (RON+MON)/2) of blended gasolines. The gas chromatographs, after correcting for minor retention shifts by internal spiking, are shown to be additive. This allows the prediction of blended product octanes directly from the gas chromatograms of the streams used.

EXPERIMENTAL

Fuels used

Eight gasoline streams were obtained from a single refinery, representing the individual components available for blending finished gasolines. These included light catalytically cracked gasoline, heavy aromatics, raffinate, light straight run gasoline, light alkylate, reformate, a toluene cut and a butanes fraction (*Table 1*). These streams were blended into 53 gasolines, regular, premium and super premium, to cover the range of typical and atypical compositions that would adequately define a reasonable data space and keep the stream variables (volume percentages) as independent as possible. A correlation matrix of the component volume percentages for both regular and combined premium/super premium fuels showed the absolute values of all pairwise correlations among fuels to be <0.7. Further, the measured pump octanes of the centre points of regular fuels and of premium fuels were close to the specification

Table 1 Blending stream properties

Stream	RON (vol%)	MON (vol%)	Aromatics (vol%)	Olefins (vol%)	Saturates (vol%)
Light cat crk	89.3	79.7	10.5	28.8	60.7
Heavy aromatics	93.9	82.8	19.3	70.5	10.2
Raffinate	66.5	65.3	3.3	3.0	93.7
Lt str run	77.0	74.0	1.8	6.2	92.0
Light alky	91.9	90.1	0.5	0.3	99.2
Reformate	108.0	95.5	94.5	0.5	5.0
Toluene fraction	108.5	107.0	99.2	0.8	0.0
Butane fraction[a]	94.0	90.0	0.0	0.0	100.0

[a] Butane values are estimates

0016-2361/90/040443-05

Table 2 Summary of octane data for 53 fuels

Octane	Min	Max	Mean	Measurement error
Research	89.4	100.7	95.7	0.25
Motor	81.1	90.0	85.2	0.30
(RON+MON)/2	85.7	94.8	90.5	0.20

values of 87.0 and 92.0, respectively. Research and motor octanes for the streams and blends were obtained in duplicate, and averaged using ASTM test procedures[8]. These data and pump octane values are summarized in *Table 2* for the blended fuels.

Gas chromatography

Gas chromatographic data were obtained on an instrument fitted with a 50 m × 0.2 mm i.d. crosslinked methyl silicone column. Column conditions included: −30°C start; heating programme of 5°C min^{-1} to 250°C; 15 min purge. A flame ionization detector was used at 350°C, while the injector was held at 250°C. Helium carrier gas at 30 psi head pressure was used with a split flow rate of 385 ml min^{-1} using a 1 μl sample size.

The raw data obtained from this chromatographic procedure were taken at the highest sensitivity and frequency of the equipment so that ≈30 000 data points per run were collected. This file was then processed with HP Lab Automation System Rev A.85 to yield a data file of 150–400 components, with each peak area attributed to a single point on the retention time scale. The data were further reduced by summing all components into retention windows of 0.1 min. To correct for minor instrumental variations, two internal standards (methylethylketone and butyl cellosolve) were used. These standards appeared in the chromatographs at relatively unoccupied retention windows and were used to make minor linear adjustments of retention times. Subsequently, these materials were subtracted from the chromatographs and the remaining areas were normalized to 100%. After this data treatment, the 666 windows from 3.5 to 70.0 min were used for subsequent calculations.

A single fuel (UP9, containing most of the refinery streams) was run 12 times over 10 days to assess g.c. repeatability. These repeat data showed that the time interval assigned to an individual component would vary by no more than ±0.1 min from run to run. This degree of precision (retention time associated with an individual component) greatly assisted in the factor analyses used to model octanes.

The effect of actual sample size and split ratio was assessed by comparing the number of components detected and the variability of total area observed for the 12 repeat runs. The number of individual components for fuel UP9 ranged from 167 to 196, and the total detector response varied from 3.2 to 4.8 million counts. Normalization to 100% and summation into 0.1 min retention windows provided data satisfactory for subsequent statistical assessment.

Data analyses

An excellent description of principle components regression (PCR) is provided by Fredericks *et al.*[7]. The data matrix D contains all the g.c. data for the model fuels. Each column is the complete chromatograph for a single fuel and each row represents the peak area associated with a fixed g.c. retention time. The product-moment matrix is:

$$Z = D'D$$

where D' is the transpose of D. If L are eigenvectors of D, then unique factors of D can be determined as:

$$F = DL$$

and

$$D = FL'$$

Thus the g.c. data matrix is factored into two matrices, F and L'. The dimensions of F are the same as D, the data matrix of chromatograms for all of the models fuels, and L' is a square matrix, with the number of rows and columns equal to the number of fuels. The columns of F and the rows of L' can often be reduced to approximately half of the number of fuels and still contain all of the information, with the deleted rows representing random noise in the g.c.. L (the principal components of D) are the transpose of L', and are used directly in multiple linear regression to predict the octanes of the model fuels. For validation fuels, the model factors F are used to determine the estimated L from g.c. data, and these values are then used to predict the octanes.

The columns of each of the matrices F and L are mutually orthogonal (independent) and the above factorization can be used for problems where the independent variables are correlated, the problem of 'collinearity'[9]. In this work the independence of the principal components L simplified the regression models and provided the means to use all of the g.c. data in octane prediction. All data processing and graphing were done using the SAS* system on an IBM 4381 computer.

RESULTS AND DISCUSSION

Assessment of any model should consider the inherent variability or noise that exists in all of the measurements. In this context both g.c. variability and octane measurement errors were estimated to provide measures of the adequacy of the models. Twelve repeat g.c. runs were made on a single fuel, and the level of g.c. variability was established. Duplicate research and motor octanes provided the estimates of octane variability and the average values for modelling. The measured octane variability is presented in *Table 2*. Direct calculation of all g.c. data from the chromatograms of the individual component streams and their volumetric blend compositions were within the variability of the repeat g.c. runs on the same fuel. Thus, predicted g.c. values calculated from components are good approximations of the measured chromatograms on the final blends. This additivity implies that factor analysis is appropriate for the g.c. data.

The g.c. data from the designed set of 53 fuels were mean centred and subjected to factor analysis. To determine how many orthogonal factors are necessary to represent the g.c. information in this fuel set, the array of 53 fuels by 666 retention windows was factored using the singular value decomposition procedure in SAS. It

* SAS is a registered trademark of SAS Institute Inc., Cary, NC, USA

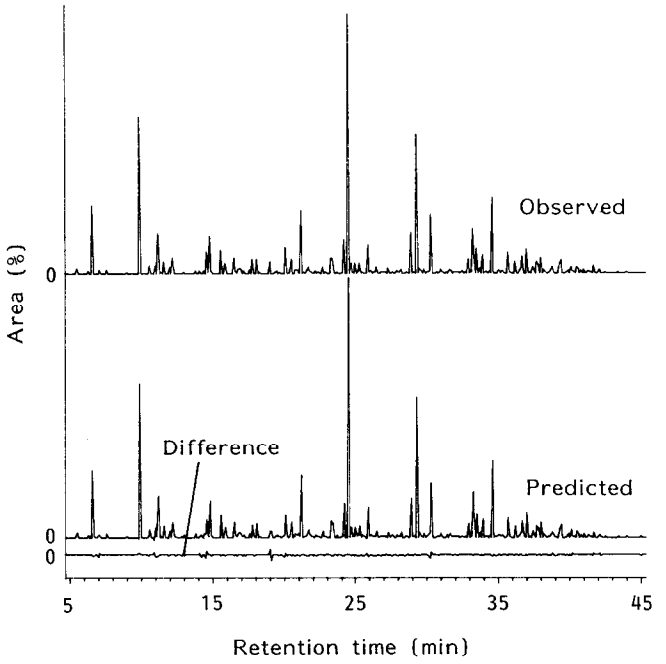

Figure 1 Gas chromatogram of fuel UP9

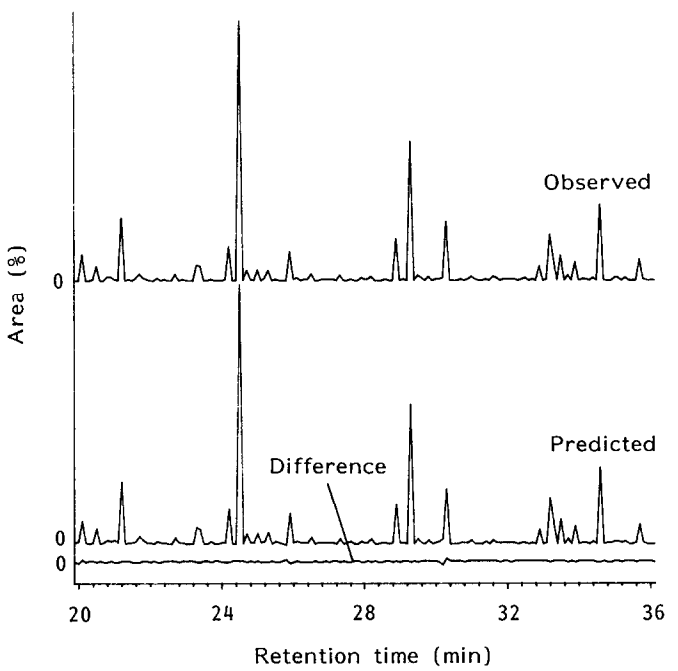

Figure 2 Segment of gas chromatogram of fuel UP9

was found that 25 factors would reproduce the gas chromatogram of fuel UP9 within the limits of variability determined by repeat g.c. runs made on this fuel. *Figure 1* graphically displays a single observed chromatogram from UP9, the calculated chromatogram from 25 orthogonal factors from the factor analysis of all 53 fuel blends, and the difference chromatogram (observed versus calculated) which is within the calculated variability observed by repeat g.c. measurements. For this reason, 25 factors were considered in modelling the g.c. data for predicting octanes. *Figure 2* displays an expanded segment of retention times and shows size and direction of the deviations encountered.

Graphical representation of the mean gas chromatogram for the 53 fuels modelled and the first six factors is shown in *Figure 3*. The symmetrical positive and negative deviations in adjacent 0.1 min retention windows are associated with peak shifting from one window to the next, due to instrumental inability to consistently assign the same retention time to a specific component. This shows how minor g.c. deviations can be accommodated by the factor analysis calculations.

Regressions of the first 25 principal components (eigenvectors) associated with the largest factors of the 53 fuels g.c. data and the average octane values overfit the data with many nonsignificant principal components. Only the significant principal components were retained in the models.

Table 3 shows the calculated range, degree of relationship, and errors associated with the models. The observed and predicted pump octane data for the 53 model fuels are shown in *Figure 4*. The error associated with predicted octanes is comparable with the error observed among the 12 repeat g.c. runs of fuel UP9 when the g.c. variability is projected into the octane space. Octane error measured in the study, reported ASTM octane measurement error, and error estimates for the same data based upon octane predictions employing an industry accepted method (generally referred to as the ethyl RT-70 linear octane blending method[10]) are included in *Table 4*. Obviously, predictability from PCR of g.c. data is good when allowing for the octane data repeatability and g.c. repeatability normally encountered.

Validation of the g.c./PCR model was accomplished by using 13 additional finished gasolines blended from

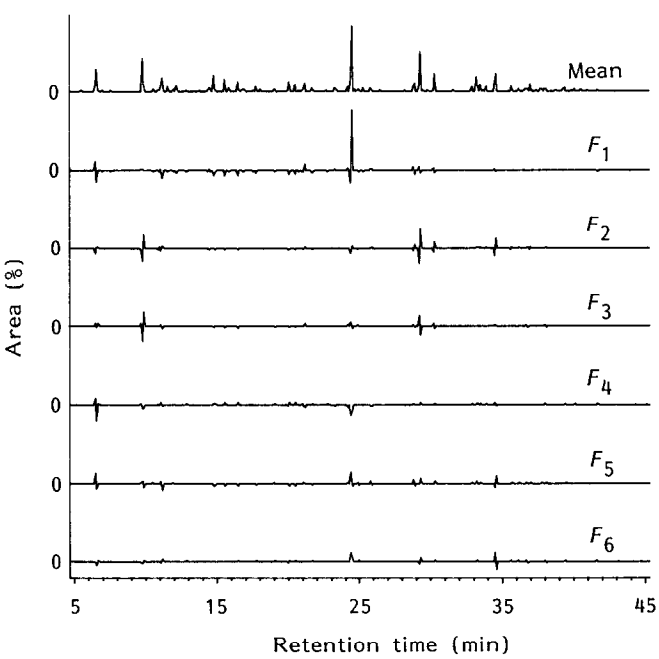

Figure 3 Mean and first six factors (factor analysis of 53 model fuels)

Table 3 Octane predictions from PCR models

Octane	Min	Max	R^2	RMS error
Research	89.6	101.3	0.98	0.48
Motor	81.2	89.8	0.98	0.32
(RON+MON)/2	85.4	95.4	0.98	0.35

Table 4 Octane variability (standard deviation of an individual measurement)

Octane	Measure[a]	Linear blend calculation[b]	PCR model	G.c.[c]	ASTM[d]	
					r	*R*
Research	0.25	0.43	0.48	0.42	0.07	0.21
Motor	0.30	0.31	0.32	0.19	0.11	0.32
(RON+MON)/2	0.20	0.27	0.29[e]	0.23	0.07	0.20

[a] Duplicate measurements
[b] Calculated using method described in Ref. 10
[c] G.c. variability projected into octane factor space
[d] Reported by ASTM, r = repeatability standard deviation; R = reproducibility standard deviation
[e] Computed from RON and MON predictions

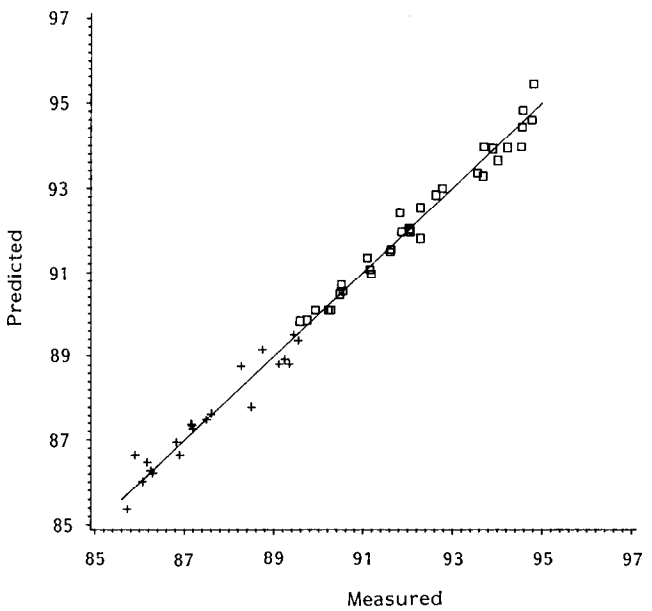

Figure 4 (RON+MON)/2 octane prediction for 53 model fuels: □, premium range fuels; +, regular range fuels

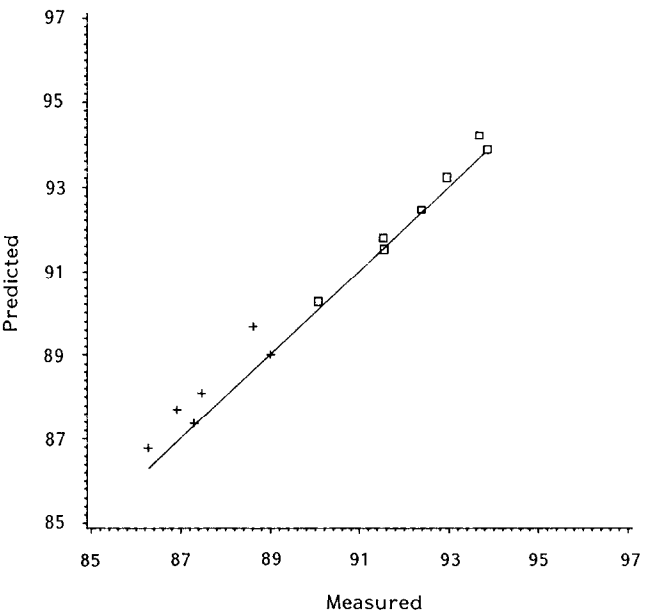

Figure 5 (RON+MON)/2 octane prediction for 13 validation fuels: □, premium range fuels; +, regular range fuels

the retained refinery streams. However, two important criteria must be met before predicting properties of an unknown sample. First, the unknown gas chromatogram must be represented by the chromatograms used in the model, and second, the magnitude of the principal components representing the sample fuel must be similar to those used in developing the prediction models. The chromatograms of the 13 validation fuels were each centred to the mean chromatogram of the model fuels, and the principal components were estimated by ordinary least squares using the 25 model factors.

The deviations (measured minus estimated peak areas) for each chromatogram and the magnitude of the estimated principal components indicated that all 13 validation blends were chromatograhically similar to the gas chromatograms used in modelling. *Figure 5* shows how well the pump octanes (RON+MON)/2 are predicted from their chromatograms, and validates this model for predicting pump octanes for the eight refinery stream system studied. The predicted values in *Figure 5* indicate a slight bias (over prediction of (RON+MON)/2). Individual RON and MON models of the premium and regular grade fuels indicate that the measured MON values are slightly low for the regular grade fuels. Substitution of other refinery streams for those modelled may or may not necessitate model recalculation, depending upon the chemical similarity between the resultant blends and the blends used in model development. However, as indicated in the validation study, the chemical similarity between an unknown and the model fuel set can be directly assessed with the g.c. model factors before predicting any fuel properties. Analogous models and validations completed for RON and MON independently were equally as good as the pump octane.

CONCLUSIONS

PCR applied to g.c. data is an ideal way of predicting the octanes of gasolines from the chemical compositions of the component streams. It is a simple method to apply since no attempt is made to interpret the g.c. data or to assign specific octane contribution factors for different compositional species. This approach should allow ready updating to include new gasoline streams or those of different chemical distributions, and has statistical controls that define the range for valid application.

REFERENCES

1 Walsh, R. P. and Mortimer, J. V. *Hydrocarbon Process.* 1971, **50**, 153

2 Anderson, P. C. V., Sharkey, J. M. and Walsh, R. P. *J. Inst. Petrol.* 1972, **58**, 83

3 Petroff, N., Boscher, Y. and Durand, J. P. *Rev. Inst. Fr. Petrol.* 1988, **43**, 259

4 Rashid, H. A., Dekran, S. B., Fakhri, N. A. and Aziz, H. J. *Fuel Sci. and Tech. Int'l.* 1989, **7**(3), 237

5 Muhl, J., Srica, V. and Jednacak, M. *Fuel* 1989, **68**, 201

6 Kelly, J. J., Barlow, C. H., Jinguji, T. M. and Callis, J. B. *Anal. Chem.* 1989, **61**, 313

7 Fredericks, P. M., Lee, J. B., Osborn, P. R. and Swinkels, D. A. J. *Applied Spectroscopy* 1985, **39**(2), 303

8 '1987 Annual Book of ASTM Standards, Volume 05.04: Test Methods for Rating Motor, Diesel and Aviation Fuels, D 2699-86 and D 2700-86', American Society for Testing and Materials, Philadelphia, PA, USA, 1987

9 Mandel, J. *The American Statistician* 1982, **36**, 15

10 Healy, W. C., Maassen, C. W. and Peterson, R. T. 'Ethyl Corporation Publication', 1959