MOTOR FUEL PROPERTY PREDICTION BY INFERENTIAL SPECTROMETRY 2. OVERCOMING LIMITATIONS

W. Marcus Trygstad
Yokogawa Corporation of America
12530 Airport Blvd.
Sugar Land, TX 77478

Randy Pell
InfoMetrix, Inc.
11807 North Creek Parkway South
Suite B-111
Bothell, WA 98011

Michael Roberto
InfoMetrix, Inc.
11807 North Creek Parkway South
Suite B-111
Bothell, WA 98011

KEYWORDS

INFERENTIAL SPECTROMETRY, CHEMOMETRICS, MULTIVARIATE MODELING, TOPOLOGICAL MODELING, GASOLINE, PROPERTY PREDICTION, NEAR INFRARED, NIR, RAMAN, NUCLEAR MAGNETIC RESONANCE

ABSTRACT

The first paper in this series, presented at the 2014 ISA AD Symposium, explored the conditions and limitations of inferential spectrometry. Data presented made the strong case that among the important factors responsible for the disparity between the promise and performance of motor fuel property predictions by inferential spectrometry is the underdetermination of sample chemistry by spectra (NIR, mid-IR, Raman, and NMR). Certainly, good models of motor fuel properties can be made for predicting motor fuel properties based on molecular spectroscopy techniques: modeling results and property predictions for unknown samples can initially appear very compelling. However, achieving continued prediction robustness may be difficult, even elusive. An important reason is that variations in gasoline composition that cause changes in properties like RON and MON do not necessarily express themselves uniquely in sample spectra. Consequently, models based on such spectra may not precisely predict changes in octane resulting from composition changes.

One solution to the problem is to perform frequent model updates to ensure that samples used in modeling correspond to the most recent refinery conditions. This approach achieves ongoing prediction accuracy at the cost of time and effort. Programs that automate the modeling are becoming available and offer the possibility to reduce the effort, but question remains about the ultimate reliability of such a “black box” approach. This second paper in the series picks up

where the first left off. Rather than jump into the topic of inferential spectrometry at the level of modeling, it explores the basis for sound chemometric analysis by contemplating implications of the "chemo" prefix through a set of vignettes based on actual applications. The prediction of motor fuel properties by inferential spectrometry seems to find its sanction predominantly in models that "look good." This paper aims instead to foster "chemical thinking" that connects the samples' chemistry (composition), their derived properties of interest, and their spectra with the modeling methodology through a sort of "information supply chain," the belief being that doing so will go far to ensure the reliable application of inferential spectrometry.

INTRODUCTION

The first paper in this series inventoried four presuppositions that underlie the application of advanced multivariate modeling methods for predicting properties of motor fuels (1):

A. The chemistry that gives rise to the property of interest expresses itself completely and uniquely in the spectral data set used for model development.

B. Chemometrics, properly applied, is capable of generating a valid model that definitively relates variance in the spectral data set to property value(s) of interest.

C. The chemistry that gives rise to the property of interest is expressed uniquely in each sample spectrum to which the chemometric model is applied.

D. Variance in the spectral data set that originates with the spectrometer or with the sample (e.g. due to temperature effects) is insignificant compared with variance originating with chemistry, or at least is ameliorated by the chemometrics.

It then went on to present strong experimental evidence that presuppositions A and C fail in the limit. Specifically, spectra of organic compounds acquired by molecular spectroscopy techniques such as mid-IR, NIR, Raman, and NMR underdetermine the chemistry of even simple hydrocarbon mixtures. That means simply that for mixtures of hydrocarbons such as those found in gasoline, changes in component concentrations do not necessary results in changes in spectra. It also goes far to explain refiners' difficulty in achieving ongoing robustness in property predictions by inferential spectrometry.

Normally, in relatively short time frames when the crude slate and the operation of conversion units vary within relatively narrow boundaries, property models can be created which are highly reliable, regardless of the spectrometric technology. However, long-term changes in crude oil and catalysts and in operating targets driven by a refinery's linear programming (LP) produce changes in the composition of blending components and in blend recipes. The resulting concentration variations in the hundreds of compounds that determine gasoline properties do not necessarily express themselves uniquely in sample spectra. Hence, chemometric models for gasoline properties generally are not robust over time, necessitating "model maintenance."

The purpose of this second paper in the series is to examine the thought and practices that either limit or promote the initial and ongoing accuracy of gasoline property models, and also the efficiency by which the desired accuracy can be achieved.

PREDICTION ROBUSTNESS: A MULTI-VARIABLE PROBLEM

Figure 1 provides an overview of the process for developing and implementing models for predicting motor fuel properties by inferential spectrometry. Due to the centrality of modeling in that process, a broad tendency exists to focus on "the modeling problem" when the quality of property predictions falls below some threshold required by engineers seeking to optimize the blending process.

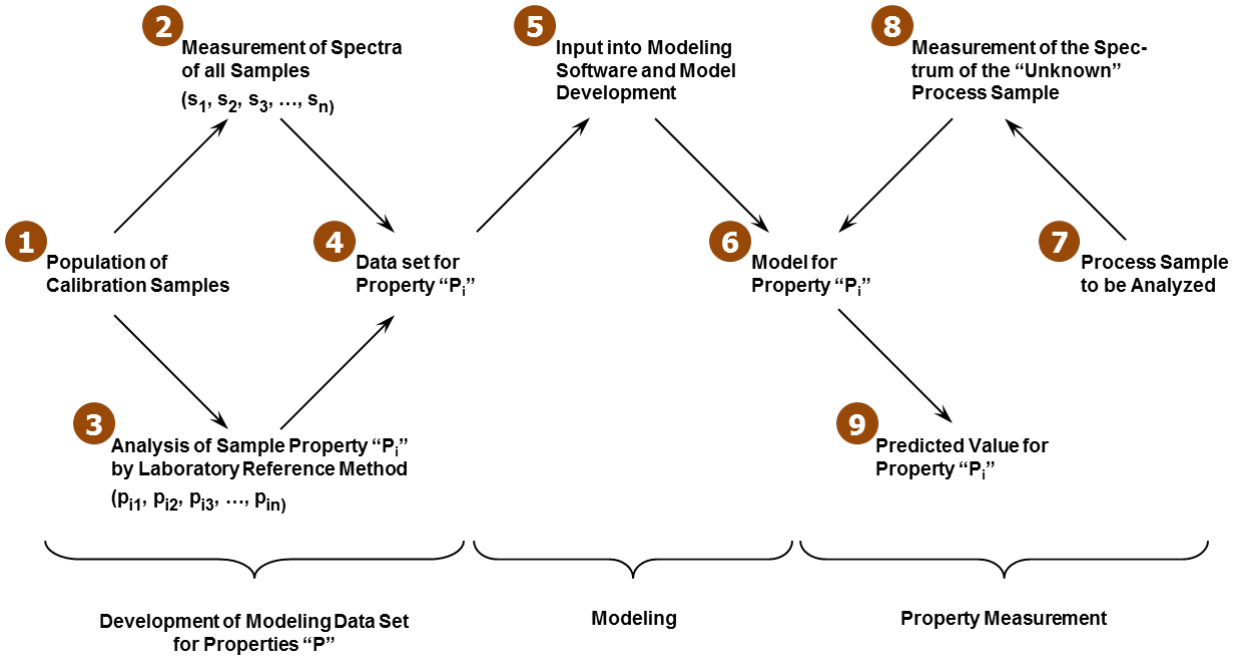


FIGURE 1. OVERVIEW OF INFERENTIAL MODEL DEVELOPMENT AND IMPLEMENTATION (PREDICTION SUPPLY CHAIN).

However, Figure 1 lays bare the fact that modeling it is not singularly important. Rather, the quality of predictions ❾ depends on eight other elements or categories of activities (2), each of which has a significant number of other associated elements or steps. Although the emphasis of the present discussion is on modeling, Figure 1 serves as an antidote to simplistic formulations that the key to inferential spectrometry resides in one particular aspect of the enterprise, e.g. a modeling algorithm or spectrometric technology. Instead, the success of inferential spectrometry demands that attention be given to all variables in the "prediction supply chain."

Overcoming Limitations. 1. When developing inferential models, recognize that the robust prediction of motor fuel properties is the outcome of not one particular activity but a comprehensive strategy that addresses each variable in the prediction supply chain. (3)

AN ALTOGETHER DIFFERENT TYPE OF PROCESS ANALYZER

The second step toward overcoming the limitations of inferential spectrometry is to recognize that "the analyzer" is not an analyzer in the conventional sense. As a rule, all other process analyzers implemented in refineries make measurements by means of some physico-chemical principal that either responds directly and selectively to the component or property of interest or elicits a response, directly and selectively. Reid vapor pressure (RVP) and zirconium oxide oxygen sensors are examples of the former while the latter includes sulfur analyzers; analysis of benzene by GC; measurement of oxygen in combustion furnaces by tunable diode laser spectrometry; and octane measurement by a knock engine. All can be regarded as "hard" measurements.

Figure 1 shows that the situation with inferential spectrometry is altogether different. Taking the example of octane, the property and the sample spectrum both are properties that are dependent on sample composition. But it cannot be said that the spectrum is an expression of octane. The common practice of referring to "the NIR analyzer" or "the Raman analyzer" obscures the fact that the spectrometer is in fact not the analyzer.

By contrast, inferential spectrometry finds strong correspondence with soft sensors applied by engineers throughout refineries to predict diverse properties of process streams. A common example is distillation properties of crude unit rundown streams. Also referred to as inferential analyzers, their implementation corresponds exactly to the process described in Figure 1. But whereas in spectrometry the independent variables (IVs) are absorbance values across a range of wavelengths or frequencies, the IVs for soft sensors are readings of pressure, temperature, and flow from field instruments installed on the process.

In every other respect, Figure 1 applies to soft sensors. Yet, engineers who refer to gasoline property predictions as being made by "the FTNIR analyzer" or "the Raman analyzer" do not refer to their soft sensors as "P-T-flow analyzers." Out of convenience, the nomenclature is unlikely to change, but this semantic digression makes the important point that the FTNIR or Raman spectrometer is no more an analyzer than an ensemble of P, T, and flow sensors. Viewed properly, the spectrometer is merely their high-fidelity counterpart, a sensor whose job it is to measure a signal that is analyzed by the real analyzer: the inferential model, which analyzes responses from the sensor to make property predictions.

Another important point is that because inferential spectrometry and soft sensors do not measure the property of interest directly and selectively, the predictions they offer have limited robustness. And the reason is the same for both: the sensor inputs into property models underdetermine the chemistry. Accepting this limitation in the case of soft sensors, engineers "tune" soft sensor outputs based on grab samples analyzed by the refinery laboratory as often as three times a day. And this is where the similarity stops: the prediction and control of the 20% distillation yield temperature for diesel rundown from the CDU is not driven by contractual or regulatory requirements. While the former has economic consequences, the latter also has legal ramifications.

Overcoming Limitations. 2. When developing inferential models, recognize that inferential spectrometry does not actually measure the property of interest.

HARD MODELS, INFERENTIAL MODELS, AND "CWODC" MODELS

On the basis of the foregoing discussion, no further distinction is necessary between measurement and prediction or between hard analyzers and those that are soft or inferential. Yet, without further qualification, it could give the mistaken impression that all model-based property predictions are equally soft, i.e. that they are grounded equally in the extent to which they relate

a) the chemical composition of the mixture
b) the spectral expression of that composition
c) the property to be modeled, whose values derive from the chemical composition.

This relationship, as depicted in Figure 2, appears to be a mere simplification of Figure 1. Yet, Table I shows that the emphasis of each is different. The question posed in Figure 2 suggests that at least two different factors can determine how effectively a property may be modeled: the distinctiveness and the directness with which the sample chemistry responsible for a property expresses itself spectroscopically.

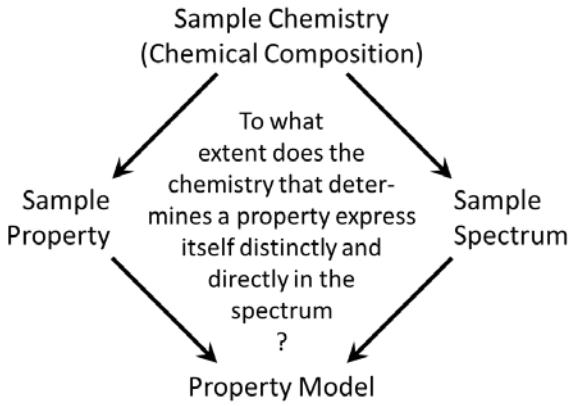


FIGURE 2. RELATIONSHIP BETWEEN A SAMPLE'S COMPOSITION AND PROPERTY MODELS (INFORMATION SUPPLY CHAIN).

TABLE I. SUPPLY CHAINS FOR MODEL-BASED SPECTROMETRY.

FIGURE 1	**FIGURE 2**
Prediction Supply Chain	**Information Supply Chain**
Emphasis: The process that must be managed and controlled to obtain model-based property predictions.	**Emphasis:** The relationship between sample chemistry, a sample property arising from that chemistry, the sample spectrum, and the desired property model.
Question Addressed: What actions and elements contribute to and must be controlled to ensure prediction accuracy?	**Question Addressed:** Do the elements of the "information supply chain" support reasoning that the model for a particular property has a basis in chemistry and spectroscopy?

Distinct Expression of Chemistry. Distinctiveness of spectral expression is the idea that the component in a mixture or the chemical basis for a property of interest exhibits a spectral band or group of bands that a) are resolved and therefore differentiable relative to bands for other components, and b) vary in intensity as a function of the property/component value. In the chemical, specialty chemical, and pharmaceutical industries, examples of such distinctiveness abound. One is organic R-OH group in polyols and nonionic surfactants: the property of interest is hydroxyl value (proportional to equivalents of OH per gram), which has a strong, well-defined absorption band in the NIR. In the petrochemical industry, BTEX (benzene, toluene, ethylbenzene, and xylenes) components in fractionation feed and rundown streams each have distinct responses in Raman spectroscopy. Examples in refining include water and HF in alkylation catalyst.

Direct Expression of Chemistry. Whereas all distinct expressions of chemistry are also direct, the opposite is not necessarily true. Octane and cetane numbers are examples: spectra of gasoline and diesel samples are an aggregate of spectra for their individual components; yet, octane and cetane as properties do not have distinct peaks as such. Nevertheless, we can reason that a strong relationship exists between

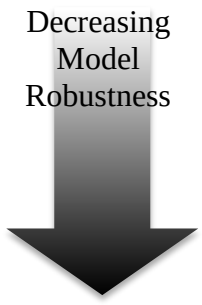
- the way chemical composition determines how a motor fuel burns in an internal combustion engine, and
- the expression of that chemistry across an appropriate spectral range of spectral responses.

Thus, while octane and cetane express themselves directly in spectra, they do not do so distinctly in the sense that there exists a peak for those properties.

Table II gives examples of the three relevant combinations of distinctiveness and directness and identifies three categories of models corresponding to each. The clear implication is that not all chemometric models are equally grounded in terms of chemistry and spectroscopy.

TABLE II. THREE DIFFERENT TYPES OF CHEMOMETRIC MODELS

EXPRESSION OF CHEMISTRY[†]	EXAMPLES	MODEL CATEGORY
Direct and Distinct	NIR: Water, HF in alkylation catalyst Raman: Benzene in a hydrocarbon mixture	Hard/Firm
Direct but Indistinct	Octane, Cetane Number	Soft (Inferential)
Indirect and Indistinct	Sulfur, RVP, (Distillation Properties, Cold Properties)	CWODC[‡] (Circumstantial)



[†] How sample composition and the property determined by it expresses itself a sample spectrum
[‡] Models that afford Correlation WithOut Direct Causation, i.e. basis for correlation is or may be unknown.

The hard/firm designation associated with Direct and Distinct conveys that although models are employed to obtain quantitative results from measured spectra, the analysis is akin to those made by other "hard" analyzers, i.e. response intensity corresponds to property value. The correspond-

ence of Direct and Distinct with soft/inferential models has been discussed above. However, a new model category is given for the Indirect and Indistinct scenario.

CWODC Models: RVP and ppm Sulfur. At best, properties listed in the last row of Table II are "very soft." In contrast to the examples of octane and cetane, unpublished work has shown that changes in concentrations of the chemical species responsible for changes in RVP in gasoline are well below spectroscopic detection limits. (The experiment is simple: 1) Measure the spectrum of a gasoline (the RVP value does not need to be known); 2) Spike the sample with an amount of butane (either n-butane or isobutane) that would result in a 0.5 pound increase in RVP; 3) Measure the spectrum of the spiked sample; and 4) Subtract the first spectrum from the second and examine the spectral residual.)

If improbable that molecular spectroscopy techniques can be applied to monitor RVP, doing so for sulfur is outright impossible. First, sulfur regulations concern elemental sulfur, and by definition molecular spectroscopy is unsuitable for elemental analysis. Some have supposed that sulfur can be measured on the basis of molecular bonds for sulfur within hydrocarbon compounds in gasoline. While true that sulfur-containing molecules have distinctive spectral signatures, they are like mere blades of grass compared with the forest of responses for the bulk hydrocarbon matrix. What's more, the sulfur is distributed between diverse molecules and functional groups.

For both RVP and sulfur, we can unequivocally state that the answer is "no" for the question posed in Table I concerning the Information Supply Chain. Someone might protest that models for distillation properties and cold properties belong in the Direct and Indistinct category with octane and cetane. They would have to explain, with reference to Figure 2, how distillation temperature expresses itself in spectral data, the problem being that molecular spectrometry is based principally variations in component concentrations.

That same protester might offer a plot such as that given in Figure 3 as evidence that the preceding discussion is idle, and that justification for a model resides not in reasoning but in correlation. Simply put, an affirmative answer to the question, "Does it work?" validates inferential spectrometry. Yet, that question begs definition of what "it works" means. One engineer offered that whether "it works" or not can be assessed very simply: Is the result used in closed loop control?

Many examples exist of correlations not undergirded by causation.

> "The slogan 'correlation does not imply causation' is meant to capture the fact that any joint probability distribution over two variables can be explained not only by causal influence of one variable on another, but also by a common cause acting on both" (4).

Expressed differently,

> "Correlation does not even imply *correlation*. That is, correlation in *the data you happen to have* (even if it happens to be "statistically significant") does not necessarily imply correlation *in the population of interest*" (5).

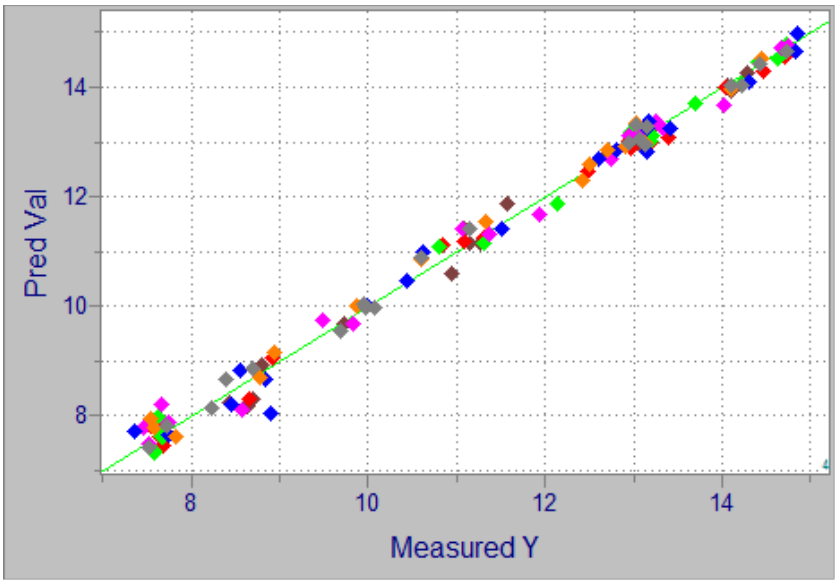


FIGURE 3. PREDICTION OF RVP IN GASOLINE BY MODEL-BASED FTNIR VERSUS MEASURED RVP.

Caution therefore must be exercised when creating models for which the information supply chain is Indirect and Indistinct, resulting in Correlation WithOut Direct Causation (CWODC). For the refinery applications of concern, the reason that correlations can be obtained should be clear. The analytical chemist tends to look at each sample as discrete, and as having its own set of associated properties. Thus, a sample of gasoline produced in a Gulf Coast refinery would have the same properties (octane, RVP, sulfur, distillation yield, percent total aromatics, percent total olefins, etc.) whether it were tested in Oklahoma or Ontario, excluding laboratory error.

However, in a given refinery, samples in a population that is the basis for chemometric models are not unrelated. They are in fact the product of a highly correlated operation, starting with the crude oil feed and ending with the blended product. That the possibility exists to obtain Correlations WithOut Direct Causation should therefore not be surprising. This explains why powerful modeling algorithms can extract correlations between spectra and properties with no spectral expression (RVP or sulfur). Such models are more appropriately referred to as Circumstantial instead of Inferential. Those who cannot be dissuaded from the have-software-will-model mindset might ask, just how indirect can the causation underlying the correlation be before a model loses its moorings in reality, and therefore its predictive robustness? A partial answer is provided by the Direct but Indistinct cases of octane and cetane number which, despite grounded in spectroscopy and chemistry nevertheless rely on deliberate updates to maintain accuracy.

Overcoming Limitations. 3. When developing inferential models, be on guard against properties whose spectral expression is Indirect and Indistinct, as this may yield Correlations WithOut Direct Causation and models with extremely limited predictive value.

ATTRIBUTES OF MODELING EFFECTIVENESS

The objective in model-based spectrometry is not modeling, but effectiveness in property prediction, where at a minimum, effectiveness is defined in terms of the characteristics described in Table III.

TABLE III. ATTRIBUTES OF MODELING EFFECTIVENESS

ATTRIBUTE	DESCRIPTION	HOW QUANTIFIED
Modeling Accuracy	The ability of a property model to account for variances in the modeled data set	Standard Error of Cross Validation (SECV) or Root-Mean-Squared SECV (RMSECV)
Initial Accuracy	Prediction accuracy evaluated by analysis of a limited number of independent samples after finalizing the model. Strictly speaking, Initial Accuracy and Modeling Accuracy should be considered as two steps in the determination of Initial Accuracy.	Root-Mean-Square Error of Prediction (RMSEP), or simply Standard Error of Prediction (SEP).
Ongoing Accuracy (Robustness)	Prediction accuracy evaluated over time following implementation of the property model for routine analysis of process samples.	RMSEP (SEP)
Efficiency	The effort level required to achieve and maintain ongoing prediction accuracy that is sufficient to support the process control objective	Cost, including: Time spent updating models (collecting data, modeling, validating results) The impact of degraded process control on product quality and operating efficiency

Chemometricians most commonly refer to Initial Accuracy and Ongoing Accuracy by the concise abbreviations SECV (or RMSECV) and SEP (or RMSEP), respectively[§]. While this paper does so as well, the identification of the attributes in Table III is important for present purposes, as it helps the non-chemometrician appreciate the distinction between the result of modeling (a model) and the result of its application to make routine property predictions of new process samples.

Embedded in Table III is the concept of three different types of validation. The first is that which occurs during modeling, with RMSECV being but of many diagnostics available to support the selection of modeling parameters and samples to be used in the final model. The second is a test of the model that is limited in time and scope whereby predictions results for a set of perhaps two dozen samples acquire over as many days are compared with values from the reference method. Initial Accuracy should be viewed as the culmination of Modeling Accuracy. Ongoing Accuracy serves two purposes: monitor that ongoing prediction performance corresponds to the Initial Accuracy; and provide samples that may be used as needed to maintain Initial Accuracy when process changes cause degradation of Ongoing Accuracy.

Refiners and vendors alike recognize that ongoing model maintenance is required, and that doing so requires vigilance, effort, and skill. Until recently, the subject of efficiency in the context of modeling and model maintenance has received little attention, much less that of cost. A paper presented at the ISA 60th Analysis Division Symposium offers hope that the subject may now be getting overdue attention (6).

Overcoming Limitations. 4. When deploying inferential models for online prediction, rigorously validate to establish initial and ongoing accuracy.

SPECTROMETRIC ERROR, REFERENCE ERROR, AND RMSEP

A common saying is that the model accuracy can only be as good as the accuracy of the reference method, the issue being that the reference method is the "lens" through which model performance is viewed. Thus, even in the limit of a hypothetical model that has no error, RMSECV or RMSEP will at best appear to have an error equal to that of the reference method.

However, consider the simple example of an experimental x-y data set that appear to relate linearly. When experimental values are plotted and a line regressed through them, the resulting linear equation is understood to be a truer representation of the relationship between x and y than any individual data pair. In other words, the regression method extracts signal (the linear model) amidst the noise of experimental error.

With multivariate models, the situation is much more complex, but the principal is the same. Applying the same reasoning used above, such chemometric modeling can be thought of a method for rejecting data set "noise" (error) to obtain high-fidelity signal in the form of a model. Indeed, the PLS modeling method is well known for its ability to reject uncorrelated noise.

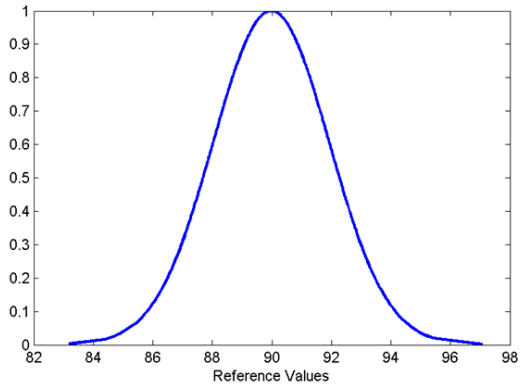
Implied in the discussion above (the model can only be as good as the reference method) is the idea that

$$\text{RMSEP} = f(\text{Reference Error, Spectrometric Error}) \quad (1)$$

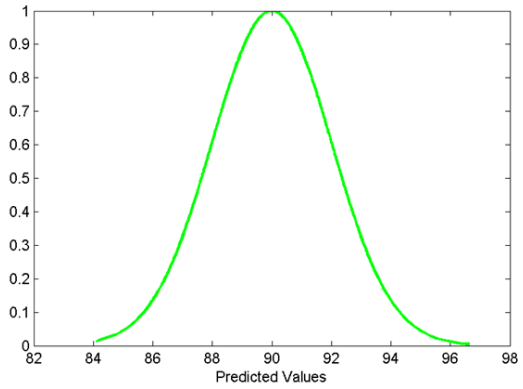
where Spectrometric Error represents the combined errors associated with the model and with the measurement of spectra by the spectrometer. Whereas the calculation of RMSEP is straightforward, the challenge is to identify an approach for teasing apart the separate error contributions from the reference and the spectrometry (the spectrometer and the model). Figures 4 and 5 represent and effort to do so. Figure 4 shows a Gaussian curve corresponding to a Reference Error of 1.94, and also five Gaussian curves corresponding to Spectrometric Errors ranging from 2.06 to zero. Given under each of the latter curves is the net RMSEP resulting from the combination of the Reference and Spectrometric Errors. That data, plus that for four additional values of Spectrometric Error are plotted in Figure 5.

The significant outcome from this data is that when Spectrometric Error has a value equal to 20% of the Reference Error, RMSEP is only 2% greater than the Reference Error (1.98 vs 1.94). These conclusions hold up when, instead of the univariate example given above, the combination

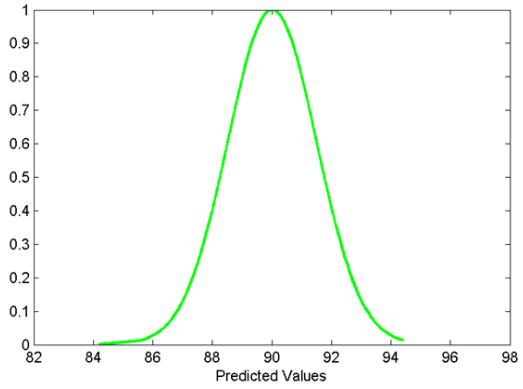
of Reference Error and Spectrometric error are evaluated in a more complicated multivariate system. Due to limitations of space, results from that analysis are not presented here, but they include a plot that is nearly identical to Figure 5.



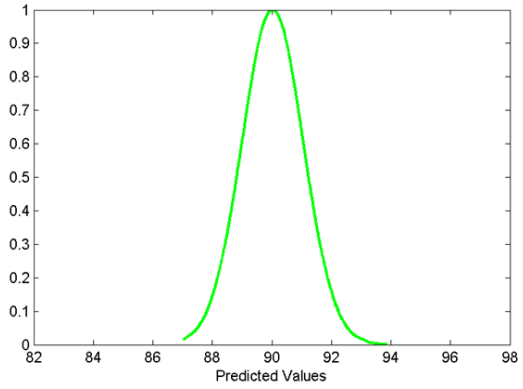
Reference Error = 1.94



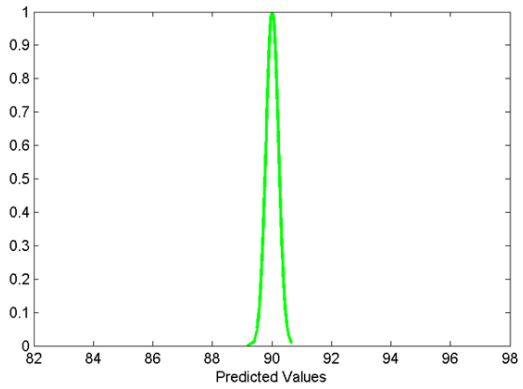
Spectrometric Error = 2.06
RMSEP = 2.83



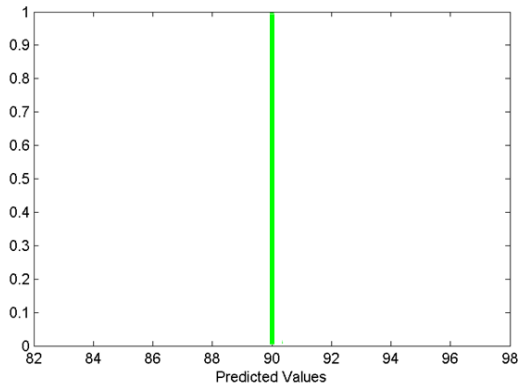
Spectrometric Error = 1.49
RMSEP = 2.45



Spectrometric Error = 0.98
RMSEP = 2.17



Spectrometric Error = 0.21
RMSEP = 1.95



Spectrometric Error = 0.00
RMSEP = 1.94

FIGURE 4. VARIATION OF RMSEP AS A FUNCTION OF SPECTROMETRIC ERROR WITH CONSTANT REFERENCE ERROR.

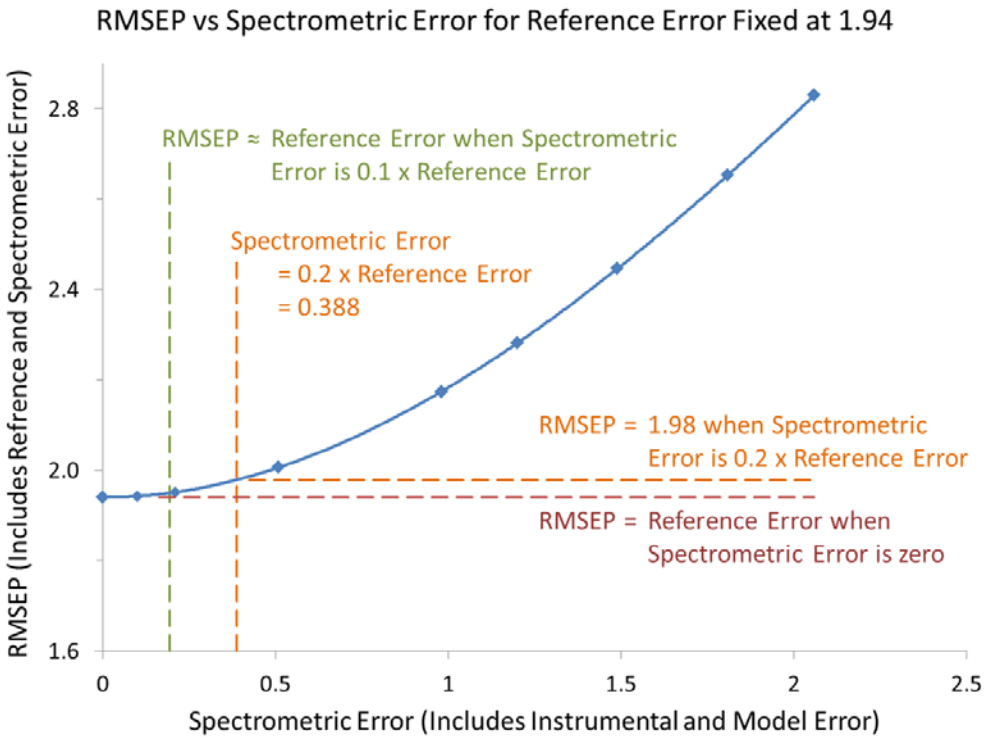


FIGURE 5. PLOT OF RMSEP VERUS SPECTROMETRIC ERROR WITH CONSTANT REFERENCE ERROR.

Overcoming Limitations. 5. When developing inferential models, RMSEP will be determined by Reference Error when Spectrometric Error ≤ 0.2 x Reference Error.

MINIMIZING MODELING ERROR

After the assertion in the preceding section, the question becomes, how can Spectrometric Error be reduced so as to be ≤ 20% of the Reference Error? This section seeks to answer that question by examining the performance that can be achieved by four modeling methods representing increasing levels of chemometric rigor.

Table IV defines those methods while Figure 6 depicts graphically the performance that can be attained by each in the analysis of octane, for which the 1-sigma Reference Error is 0.204. Method D has been shown to consistently deliver RMSEP values within a few percent of the Reference Error, the conclusion being that the Spectrometric Error is ≤ 20% of the Reference Error. RMSEP values for Methods A, B, and C are based on actual RMSEP values for those methods compared with Method D. With Reference Error being constant across that range of methods, it was possible to estimate the Spectrometric Error according to Equation (2).

$$\text{Spectrometric Error} = \sqrt{(RMSEP)^2 - (Reference\ Error)^2} \qquad (2)$$

Rigorous analysis of error propagation through a multivariate algorithm is extremely difficult if not impossible. Equation (2) nevertheless provides a useful means for estimating Spectrometric Error in terms of the known values for RMSEP and Reference Error. The results are consistent with what has been demonstrated empirically for univariate and multivariate systems.

TABLE IV. DEFINITION OF FOUR MODELING METHODS

Method A. Typical Practices. Typical for individuals trained in the procedure based application of chemometric software (PLS*) but lacking broad experience and knowledge of fundamentals
Method B. Common Practices. Typical for scientists with solid understanding of spectroscopy and chemometrics who apply standard modeling tools like PLS appropriately
Method C. Enhanced Practices. Adds a RobustPLS algorithm to enable optimum selection of samples for use in modeling (8)
Method D. Best Practices. Adds modeling tools including spectral TuneUp™ (9)

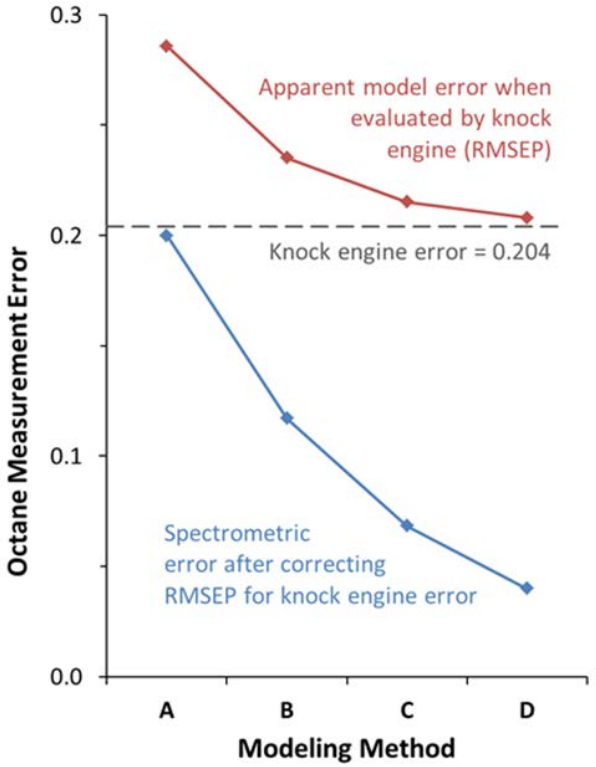


FIGURE 6. APPARENT VERSUS ACTUAL OCTANE PREDICTION ERROR FOR FOUR MODELING METHODS

The improvements in RMSEP for Methods C and D are small compared with Method B (red data in Figure 6); a chemometrician laboring to achieve octane predictions with the highest possible accuracy might conclude that on the basis of RMSEP that the benefits of Methods C and D are not compelling. However, Figure 6 reveals a dramatic decrease in the underlying Spectrometric Error (blue data), which is due to the manner in which errors combine through Equation (2). Thus, RMSEP belies the magnitude of the actual improvement in prediction performance afforded by Methods C and D compared with Method B.

Methods A and B employ conventional modeling based on PLS whereas Method C adds RobustPLS (7) (8), a licensed technology for automating the objective identification of spectra that should NOT be retained for calibration. Once optimized for a particular instrument and reference method, it can rapidly define the calibration set with results comparable with, even superior to, analysis by an expert. RobustPLS helps avoid two scenarios that confound outlier identification by typical visual means:

- Masking - when bad data points are hidden by other bad data points and thus appear to be good points
- Swamping - when bad data points make good data points look like bad data points

Method D represents a further enhancement of Method C by the application of proprietary pre-processing techniques and a further tuning of the application of PLS through a well-known technique called locally-weighted regression (LWR), which is a form of topological modeling. The TuneUp™ spectral alignment algorithm is a particularly valuable tool for ensuring maximum integrity of spectral registration when models developed on a laboratory spectrometer are implemented for measurements with an online FTNIR.

Although FTNIR is generally considered highly reproducible, deviations in spectral registration smaller than a few tenths of a wavenumber have been shown to have chemometric consequences, and that preprocessing with TuneUp™ improves spectral sensitivity to chemical changes (9). The ability of TuneUp™ to align spectral data sets with a precision better than 0.1 cm^{-1} exceeds the spectral reproducibility specification used by FTNIR manufacturers for testing and controlling spectrometer quality in production.

Overcoming Limitations. 6. The application of advanced chemometric methodologies yields improvements in RMSEP that appear relatively inconsequential, but which are due to dramatic reductions in underlying Spectrometric Error to levels below 20% of the Reference Error.

REFERENCES

1. Trygstad, W. Marcus, Horgen, Dana, "Motor Fuel Property Prediction by Inferential Spectrometry: Understanding Conditions and Limitations", AD.14.10.0 in Proceedings of the ISA 59th Analysis Division Symposium, Baton Rouge, Louisiana, May 4 – 8, 2014.

2. Ibid.

3. "Overcoming Limitations" presented in this paper represent a high-level companion to "The Six Habits of an Effective Chemometrician" defined by Beebe, Kenneth R., Pell, Randy J., and Seasholtz, Mary Beth, Chemometrics, A Practical Guide, Copyright © 1998, John Wiley & Sons, Inc.

4. Ried, Katja , *et al*., A quantum advantage for inferring causal structure, Nature Physics (2015).

5. http://andrewgelman.com/2014/08/04/correlation-even-imply-correlation/

6. Pell, Randy J., Roberto, Michael F., Ramos, Scott, and Rohrback, Brian G., "Re-Engineering Calibration in Optical Spectroscopy", AD.15.01.02 in Proceedings of the ISA 60th Analysis Division Symposium, Galveston, Texas, April 26 – 30, 2015.

7. Rousseeuw, Peter J., "Tutorial to robust statistics", Journal of Chemometrics, 1990, 1.

8. Hubert, Mia and Vanden Branden, Karlien, "Robust methods for partial least squares regression", Journal of Chemometrics, 17, 2003, 537-549.

9. Trygstad, W. Marcus, Pell, Randy, “Spectral TuneUp for Improved Motor Fuel Analysis”, 28th International Forum and Exhibition, Process Analytical Technology – IFPAC, Washington, D.C., January 24, 2014.

§ Root-Mean-Squared Error of Cross Validation, or RMSECV, is a measure of agreement between reference values for calibration samples and values predicted for those same samples during model development. Sometimes referred to simply as SECV, it is obtained by iteratively removing calibration samples from the data set and predicting them as unknowns, hence the term “cross validation.” The calculation of Root-Mean-Squared Error of Prediction is the same as that for RMSECV. However, instead of being based on calibration samples, it compares predicted and reference values for an independent sample set. Consequently, RMSECV provides an optimistic assessment of prediction error while that provided by RMSEP is more objective and tends to be higher than RMSECV. However, the extent to which that is the case depends largely on the proper gathering of calibration samples, the proper selection of model “inliers,” and on avoiding over-fitting of the modeling data.