Impact of near infrared signal quality on chemometric model predictions

Noémie Caillol[1]*, François Wahl[1], Nathalie Szydlowski-Schildknecht[1], Bertrand Lecointe[2]

[1]IFP Énergies Nouvelles, Établissement de Lyon, Rond-point de l'échangeur de Solaize, BP 3, 69360 Solaize
[2]IFP Énergies Nouvelles, Établissement de Rueil
*Corresponding author: noemie.caillol@ifpen.fr

Introduction

Real-time access to fuel property characterisation is needed for online process control (process analytical technology; PAT), or onboard vehicles for optimised engine settings. Near infrared (NIR) based micro-opto-electro-mechanical systems (MOEMS) are being developed that will be smaller and cheaper than classic laboratory equipment. Miniaturisation unfortunately affects signal quality. The aim of this study is to assess the impact of signal qualities such as resolution, sampling interval, signal to noise ratio (S/N) and spectral window width and position on model predictability. Our ultimate objective is to define the most appropriate hardware for end-users, which will reflect the need for precision. In our present case, the aim was to determine the NIR signal quality required for predicting two properties of diesel fuel.

Materials and Methods

Previous work has established models for predicting cetane number[1] and aromatic content[2] in diesel fuel. These models were developed using a laboratory NIR spectrometer (ABB-Bomem) on a wide sample base. Spectral degradations were subsequently simulated in order to determine the impact that low quality spectra can have on calibration models.

Initial reference models

Initial models were built from more than 250 samples of diesel fuel. Spectra were recorded on the aforementioned NIR spectrometer at a controlled temperature of 27.5°C with a 4 cm^{-1} resolution. Reference values of investigated properties were obtained from standard reference methods.[3,4] Models were developed using partial least squares (PLS) after classic spectral pre-treatment by derivative or baseline correction between 4450 and 9000 cm^{-1}. These initial models have been in use for several years and are used as reference models for the present work as they are considered to be developed from 'ideal spectra'.

Mathematically simulated degradations

All spectral degradations were simulated using Matlab, and a general description for spectral window, S/N and resolution reductions is given below. PLS models based on these degraded spectra were then compared in terms of root mean square error of calibration and prediction (RMSEC and RMSEP).

Spectral window reduction

The spectral window was programmed by adapting the spectral range according to the central wavenumber of the MOEMS device. Compared to the 4450 – 9000 cm^{-1} range usually used for fuel models, only a continuous window of about 25% of this range had to be selected.

Signal-to-noise ratio

$$S/N = \frac{\text{signal maximum amplitude}}{\text{noise standard deviation}}$$

S/N was experimentally calculated as the division of the maximum absorbance in a diesel fuel spectrum within the considered spectral range, by the standard deviation for noise measured in a region of no absorbance in the spectra (6365–6390 cm^{-1}). The S/N obtained for 1 scan from the laboratory NIR spectrometer was 600, which increased to 5000 for a spectrum with 100 scans.

The mathematical degradation of S/N consists of adding random noise to the reference spectrum, where the noise follows a normal distribution with a standard deviation termed σ_{add}. The noise is obtained from $\sigma_{sensor}^2 = \sigma_{ref}^2 + \sigma_{add}^2$, where σ_{sensor} and σ_{ref} are the standard deviation of noise for the sensor to simulate, and for the reference spectrum respectively. An illustration of a noised spectrum is shown Figure 1.

Reference paper as:
N. Caillol, F. Wahl, N. Szydlowski-Schildknecht, B. Lecointe (2012).Impact of near infrared signal quality on model predictions, in: Proceedings of the 15th International Conference on Near Infrared Spectroscopy,
Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 207-210.

Experimental validation of this simulated degradation was done by comparing a 1 scan Bomem spectrum to a simulated S/N of 600 using a 100 scan spectrum as reference. The spectral difference of the two was less than the tolerated repeatability difference of 5 mAbs.

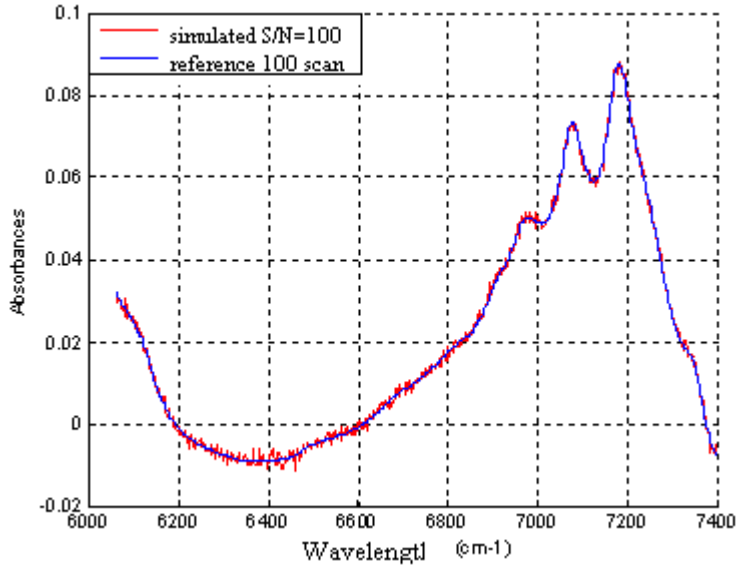
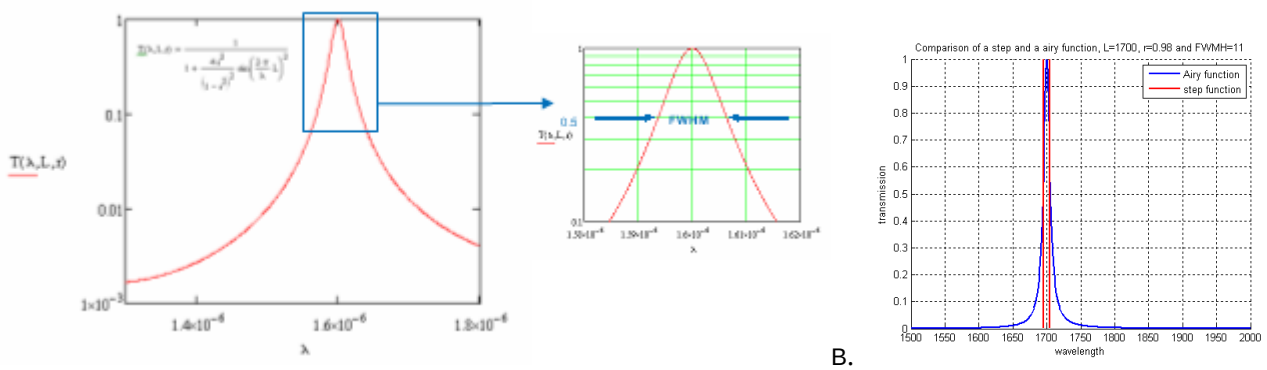


Figure 1. Zoom of diesel reference spectrum in blue (mean of a 100 scans, S/N = 5000) and the corresponding simulated spectrum with a S/N of 100.

Resolution

Resolution describes the capacity of a detecting system to distinguish two objects. In NIR instrumentation, it is often characterised from the full width at half maximum peak height (FWHM) of the transmission curve of the analyser. For instance, the transmission of a Fabry-Perot interferometer can be described by an Airy function as illustrated in figure 2. From the experimental transmission curve of a system, the measured FWHM gives the expected resolution.



A. B.

Figure 2. A) Shape of an Airy transmission function with a zoom to illustrate its FWHM. B) Comparison of an Airy function and step function of an equivalent FWHM.

As a first approximation for simulating resolution levels, the transmission function of the instrument was modelled as a simple step function of FWHM width. The mathematical degradation of resolution consisted of smoothing the reference spectrum by taking the mean of absorbances surrounding every wavelength within a window of FWHM width (in wavelength units). Experimental validation of this simulated degradation was positively done by comparing a 64 cm^{-1} resolved Bomem spectrum to a simulated 64 cm^{-1} resolved spectrum using the 4 cm^{-1} resolved spectrum as reference. With resolution, the sampling interval could also be adapted by selecting only one point out of every given sampling step (x). Spectra for which x was smaller than the FWHM could be considered "over-sampled".

Reference paper as:
N. Caillol, F. Wahl, N. Szydlowski-Schildknecht, B. Lecointe (2012).Impact of near infrared signal quality on model predictions, in:
Proceedings of the 15th International Conference on Near Infrared Spectroscopy,
Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 207-210.

Results and Discussion

Impacts of different NIR acquisition parameters on the spectra are illustrated in Figure 3, which shows a laboratory reference spectrum used for predicting diesel properties, and a corresponding "sensor-like" spectrum. The different characteristics of these spectra are detailed in Table 1.

Table 1. Laboratory spectrometer and sensor performances

	Lab NIR Spectrometer	Sensor
Spectral zone	4000–12 000 cm^{-1}	1350–1650 nm
Optical path	2 mm	1 cm
Resolution	4 cm^{-1} (4150 points)	~ 10 nm
Sampling points	696 points in 6060–7400 cm^{-1} 50000	30 points
S/N	~5000 in 6060–7400 cm^{-1}	100

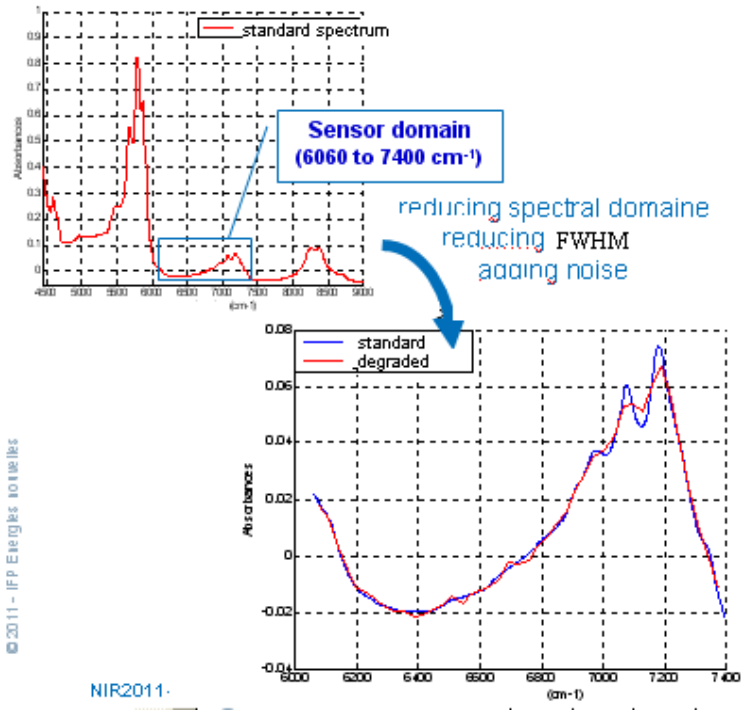


Figure 3. Transformation of the reference spectrum to a given simulated sensor-like spectrum.

Impacts of NIR acquisition parameters on spectra were evaluated using the fuel property prediction models; results are given in Table 2. Each NIR quality was studied separately and gave a good understanding of NIR model robustness and sensitivity. From the initial database available for each property, a quarter of the samples were selected via a Kennard & Stone algorithm for prediction purposes and models were calibrated on the remaining three quarters of the dataset. To compare our models to the reference method we usually calculated the percentage of samples predicted within the confidence interval of the reference method (% in CI). In the case of the aromatic content model, an interval of ±2% was selected.

Table 2. Model performance according to spectral signal quality.

		Cetane model			Aromatics model		
	Spectral Domain (cm^{-1})	RMSEC	RMSEP	% in CI	RMSEC	RMSEP	% in ±2%
IFPEN ref	4900–9000	1.84	1.89	96.4	0.75	0.79	96.8
Reduced zone	6060–7400	2.66	2.97	78.3	2.95	3.18	58
Degraded S/N	4900–9000	1.78*	1.93	93.9	0.30*	1.01	77.7
Degraded resolution	4900–9000	1.87	1.85	95.2	0.97	1.17	83.0
Entirely degraded	6060–7400	3.46	3.93	66.3	6.60	7.43	12.8

Model results suggest that the spectral zone intended for this sensor is useful for cetane number prediction (RMSEP increased from 1.89 to 2.97 but is sufficient considering the precision requested). Models indicated that the sensor is not appropriate for predicting the concentration of aromatics. The selected spectral region did not contain enough chemical information to describe the aromatics content of diesel fuels.

S/N and resolution did not significantly affect model performance when larger spectral regions were used. Combining a reduced S/N or resolution with a reduced spectral range, however, substantially lowered model performance. Indeed, much of the information in a NIR spectrum is redundant (i.e. often 90% of the variance of our diesel database is expressed along the first principal component of a PCA), So even if this information is noisy or poorly resolved, due to the fact that it is often present along the spectrum, PLS models seem to be able to compensate for its low quality. However when the information is not redundant enough, PLS is not so effective.

Reference paper as:
N. Caillol, F. Wahl, N. Szydlowski-Schildknecht, B. Lecointe (2012).Impact of near infrared signal quality on model predictions, in:
Proceedings of the 15th International Conference on Near Infrared Spectroscopy,
Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 207-210.

As a result from these simulations, we were able to establish that the sensor as it was first intended would not have satisfied the required precision for the targeted application. However, knowing the sensor potential characteristics, further simulations have allowed us to determine the optimal configuration. Compromises between number of sampling points and noise had to be reached, as well as between spectral window width and resolution, depending on the property being predicted.

Overall, simulations showed that a reduction of the spectral window width had the strongest impact on the prediction quality. This impact could be minimised if the window location along the spectral range was well selected. For instance, on a given sensor configuration, the RMSEC for aromatic content dropped from 2% with the full spectra to only 4% with a spectrum width reduced to 10%. The spectral zone to be preferred when modelling most of the diesel properties is the main fully resolved absorbing region of the spectrum (see Figure 4). Its bands correspond mainly to the first harmonic of CH_2 and CH_3 stretching vibrations. The initial domain suggested contained the combination bands: $2\nu(CH_2) + \delta(C\text{-}H)$.

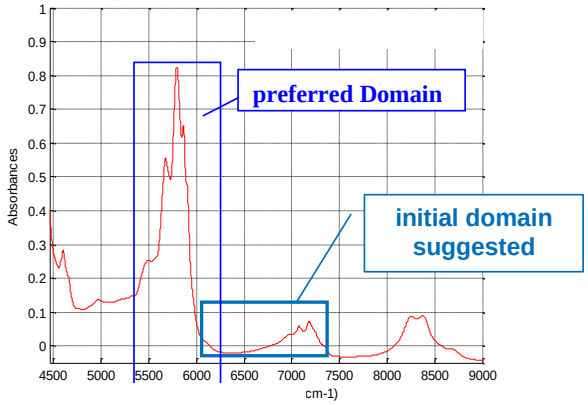


Figure 4. Selection of the spectral region.

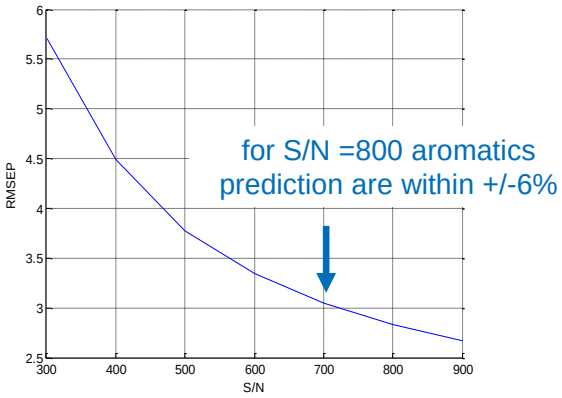


Figure 5. RMSEP for aromatics model according to S/N.

Our study showed that S/N could dramatically affect predictability of models. Figure 5 shows how the RMSEP for the aromatics model evolved according to the S/N ratio simulated on the spectra. From this information, it is possible to determine the minimum S/N required to satisfy a specific target for prediction performances. A minimum S/N of 800 was required for the study presented here. Spectral acquisition strategies can be implemented to reach this minimum S/N if the hardware allows it (i.e. multiple scanning to reduce noise, wider optical path to maximise signal). From simulation studies based on existing reference data, spectra and models, it is possible to evaluate the potential of new spectrometer devices and optimise their configuration to our needs.

Conclusion

A new approach is proposed for determining NIR analyser performances according to target precisions of models. This approach is based on the comparison of model performances developed from a unique set of spectra on which signal degradations are simulated. The impact of signal qualities(resolution, sampling, signal to noise ratio and spectral window width and position) on model predictions were assessed. Prediction was very dependent on NIR signal quality, even though all parameters did not affect models to the same degree. Spectral window selection was a key parameter since chemometric models will not be able to predict anything if the chemical information is not present in the spectra.

Acknowledgements

The authors acknowledge the Eurepides Council for funding the European project: IQFuel (Integrated sensor for determining the Quality of Fuel) and all the partners of the project.

References

1. S. Aji, N. Schildknecht-Szydlowski and A. Faraj, *Oil Gas Sci. Technol. – Rev. IFP*, **59** 303-321 (2004).
2. N. Szydlowski-Zanier, M. Berger, F.Whal, D. Guillaume, *J. Near Infrared Spectrosc*. **11**, 83-95 (2003).
3. ASTM D613: Standard Test Method for Cetane Number of Diesel Fuel Oil.
4. IFP9409: Petroleum products, mono-, di-, polyaromatic content by UV absorption spectrometry.

Reference paper as:
N. Caillol, F. Wahl, N. Szydlowski-Schildknecht, B. Lecointe (2012).Impact of near infrared signal quality on model predictions, in:
Proceedings of the 15th International Conference on Near Infrared Spectroscopy,
Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 207-210.