Detecting trace gases with optical emission spectroscopy and supervised machine learning

Paul Maguire, Jordan Vincent, Tahereh Shah Mansouri, Omar Nibouche, Hui Wang
University of Ulster

Detecting trace constituents of gases via atomic emission spectroscopy has a long history. The advent of low temperature atmospheric pressure plasmas opens up the possibility of using emission spectroscopy for portable and/or low cost detection in a diverse range of applications from industrial leak detection to volatiles and clinical breath analysis. However the nature and quality of the spectra from these plasmas, and the use of low resolution portable spectrometers, prevents the direct and reliable identification of trace constituents. Therefore we are exploring the use of machine learning techniques to develop predictive identification models based on optical emission training samples. Our original work involving unsupervised principal component analysis indicated significant cluster separation even at sub-ppm levels of e.g. NO impurity gases. Recently we have investigated supervised learning using Partial Least Squares Discriminant Analysis (PLS-DA) using a dataset of He-CH4 spectra where the CH4 concentration varies from 0 – 100 ppm. Methane is a representative hydrocarbon gas found in a number of fields from breath analysis to natural gas production and research is ongoing into accurate environmental CH4 detectors in the ppm range.

The spectra were obtained from an RF plasma formed in a quartz capillary between two exterior ring electrodes. The capillary outlet was a large distance (~100 cm) from the plasma to minimise atmospheric impurity back-diffusion and the system was initially conditioned to remove background impurities, over 21 days, using a 100% He plasma and exterior IR heating while monitoring spectral impurity bands. Spectra were obtained using a Ocean Optics HR4000CG-UV-NIR spectrometer in the wavelength range 194 – 1122 nm (interval 0.25 nm), with a slit width of 5 μm and a minimum optical resolution >0.5 nm. Data is collected in a matrix of 3648 variables (wavelengths) and 523 samples in columns, which form 9 CH4 concentration categories (0, 1, 2, 4, 6, 12, 23, 77, 100 ppm). The time duration for recording this data is 9480s. Spectral features corresponding to He, carbon, hydrogen and impurities (N, O, OH/H2O) were observed. No peak, except possibly near 516nm (C2 Swan bands), can be assigned unambiguously to any particular species. The dominant peaks for 0% CH4 were at 587.95 nm and 707.08 nm, which can be assigned to He I (587.559 ... 587.596 and 706.5) and their intensity varied by ~26% (std. dev). On introduction of CH4, the intensity of these peaks remains constant (within 1 std. dev) up to ~23 ppm and falls thereafter. Discrimination of C I (587.734, 588.95, 706.58, 707.1 ... 707.648) and C II peaks (587.95 ... 588.97, 706.36) is not possible due to the spectrometer resolution. A small peak at 778.5 nm, possibly O I (777.54), appeared in all spectra with almost constant intensity while other peaks varied arbitrarily with increasing CH4 concentration. Analysis of spectra using MassiveOES and Specair provides some indicative information but the low spectrometer resolution prevented more detailed modelling.

Predictive models were generated by PLS-DA by splitting the data samples into various training and test sets. Two general approaches were explored, namely (i) 2 class and (ii) 8 class. In the former, a threshold concentration was set and the model use to predict whether an unknown sample was above or below this threshold. For a threshold value of 2 ppm CH_4, the model accuracy was > 95% with < 10 latent variables (LV). In the 8 class model, the accuracy reached > 90% (< 10 LV) after pre-processing of spectra to include autoscaling, smoothing and baseline correction. From a computation perspective, our OES spectra represent high dimensionality collinear data with a temporal drift and low resolution. This presents a serious challenge to developing robust machine learning algorithms. It is necessary to gain insight into how algorithms function and their sensitivity to OES features. Variable Importance in Projection (VIP) is a technique to determine the relative importance of variables to predictive accuracy. In the 8-class model the most important peaks are those with the highest intensities. In the 2-class model, a similar picture is observed. However VIP suggests the most important peak is at 336.4 nm. The peak intensity is low but increases with CH_4 concentration. Species bands near 336 nm include O_2 and N_2 impurities as well as C II. The use of VIP and other feature selection methods will allow us to reduce the model complexity by removing redundancy and limiting overfitting. We have also explored variable reduction, in the 8-class model, by splitting spectra into up to 10 arbitrary wavelength ranges and building models based on each spectral subrange. Although the overall accuracy is reduced, we observe improved performance for subranges in the near IR region, where the number of peaks is limited. For the future development of this technique, we need to develop more targeted algorithms and efficient feature selection protocols in order to tackle samples containing more complex molecular mixtures, including air and H_2O, as well as further reduce the spectrometer resolution.