Application Note

SPORIAN®
MICROSYSTEMS, INC

In-line Real-time Raman Spectroscopy for Wastewater Monitoring

Authors

Karen Keiter
Duc Nguyen
Kevin F. Harsh
Sporian Microsystems Inc.

Introduction

Monitoring wastewater quality parameters is essential for a wide range of current and emerging industries in order to meet requirements of local governing agencies, and the Water and Wastewater Treatment Market is expected to reach a value of $242.6 billion by 2027[1]. Water quality in the U.S. is legislated through the Clean Water Act to reduce the discharge of pollutants into rivers, lakes, or oceans. Industrial wastewaters are often cited as the main sources of inorganic heavy metal pollutants (such as arsenic, chromium, and selenium) which are toxic and non-biodegradable [2]. These wastewaters are generated from industries such as electric and nuclear power plants, pulp and paper, iron and steel, mines and quarries, agricultural and food operations, and chemical manufacturing [3]. Depending on the origin and discharge of the wastewaters, the precise concentrations and regulatory chemical constituents of concern will vary.

In situ, real-time wastewater composition monitoring would be of significant value for improving process control, efficiency, and costs of operation. The method of ion chromatography currently plays a large role in wastewater analysis, but it is burdened by sample preparation requirements, lab measurement times, overlapping detection regimes, and the generation of additional chemical wastes [4]. There are currently few technical options for a robust monitoring system, with simple operation, that allow for detection with speciation and quantification of multiple wastewater constituents in real time.

Raman spectroscopy is a powerful analytical technique that quickly gives highly specific information for the analysis of chemical compounds in a non-destructive manner. Raman-active species exhibit spectra with distinct peaks and provide "fingerprint" information on the vibrational transitions within a molecule by uniquely characterizing sample volumes in bulk. Raman systems are easily portable and signals can be transmitted by optical fibers over long distances for remote analysis. These features make Raman spectroscopy ideal for implementing a robust, automated molecular identification system.

Ordinarily, dealing with a broad spectral range of multiple Raman-active species requires significant expert post-processing and interpretation. However, systems that deploy automated machine learning (ML) models and algorithms can significantly streamline Raman data processing and vastly improve the quantity and quality of information from available data. ML, a sub-category of artificial intelligence, is a data-driven approach to analysis, where instead of fitting a physics-based theory to the data output, an algorithm is used to leverage relationships within the data to generate correlating models. ML is therefore suitable for situations where the system outputs are unknown, hard to understand, or time-consuming to process by a human analyst. ML models can learn from spectral data, identify patterns, and be able to make decisions with minimal human intervention. In this case, the models rapidly process large amounts of complex spectral data in order to classify chemical compounds and determine their concentrations.

This application note describes the classification and quantification of thirteen chemical wastewater constituents with the use of a real-time Raman spectroscopy monitoring instrument. ML models are tested on training spectral data and then applied to a wastewater sample collected from a fossil fuel plant flue gas desulfurization (FGD) effluent stream.

Methods

Data presented here were collected using a Sporian Microsystems SpecIQ™ Raman Fluid Composition Monitoring System. Key system specifications are shown in Table 1. The system is designed to operate in one of two possible modes: (1) as an autonomous monitoring device that measures and applies machine-learning-based algorithms to provide classification of chemical constituents and quantified concentrations to higher level control systems, and (2) as a user-operated instrument with specific feature/function control options. In this example, we use the latter mode to generate training data to allow for either mode of operation in a wastewater system.

Table 1: Key SpecIQ™ Specifications

Spec/Feature	Unit
Excitation Wavelength	532 nm
Wavenumber Range (Shift)	100-5400 cm^{-1}
Resolution	6 cm^{-1}
Measurement Temp. Used	20°C
pH Range Used	5-7
User interaction/sample prep	None
Operational Pressure range	<50 psi
Liquid measurement volume	~4 cm^3
Onboard data processing?	Yes
Design for industrial/rugged use?	Yes
Communications/Interface	Ethernet/SCADA
Supported Power	110AC/28DC

The following species were chosen for examination via Raman spectroscopy as examples of soluble species commonly found in industrial wastewaters, such as those from fossil fuel power plants flue gas desulfurization (FGD) processes or mining waste treatment.

- Selenate (SeO_4^{2-})
- Selenite(SeO_3^{2-})
- Selenocyanate (SeCN)
- Boric acid ($B(OH)_3$)
- Hypochlorous acid (HOCl)
- Nitrate (NO_3^-)
- Nitrite (NO_2^-)
- Sulfate (SO_4^{2-})
- Bicarbonate (HCO_3^-)
- Orthophosphate ($H_2PO_4^{2-}$)
- Perchlorate (ClO_4^-)
- Chlorate (ClO_3^-)
- Sulfite (SO_3^{2-})
- Arsenate ($H_2AsO_4^-$)
- Chromate(CrO_4^{2-})

While the process of Raman spectroscopy requires no sample preparation, standard concentrations of each species were prepared gravimetrically from commercially available reagents (Sigma Aldrich) dissolved in UV grade water. This included serial dilutions of each constituent at concentrations bounding those nominally encountered in wastewater treatment processes. During measurement, samples were contained in a Teflon-lined sample cell and measured from below through a sapphire window.

All spectral data provided were processed and analyzed by performing a sequence of preprocessing, subsequent dimension reduction, and ML classification and regression. Preprocessing included the stabilization of signal fluctuations based on laser output and the removal of background signal. Principal component analysis (PCA) was performed as a dimensional reduction and to find the optimal variance in the data. A supervised ML model (via Scikit-learn) was implemented to train on the known dataset, analyzing their relationships. Scikit-learn's Python library was used for the entire ML process [5]. Scikit-learn is an open-source ML library that supports supervised and unsupervised learning. For simple identification of material, a classification model was then used that produced a confidence interval (CI) in the classification of the material spectra.

Results

Figure 1 shows overlapping preprocessed Raman spectra of chemical constituents of interest in slightly acidic FGD wastewaters. All samples were measured individually, at high concentrations (100 mM), and plotted as the average of 20 scans. As expected, these aqueous anions often have multiple Raman peaks, of wildly different intensities, with subtle peak shifts depending on the measurement conditions (pH, temperature, ionic strength). To show that the simultaneous detection of different constituents can be made, regardless of differences in concentration, a large training dataset was first generated of each constituent on its own. Continuing work is still being applied to mixture training data.

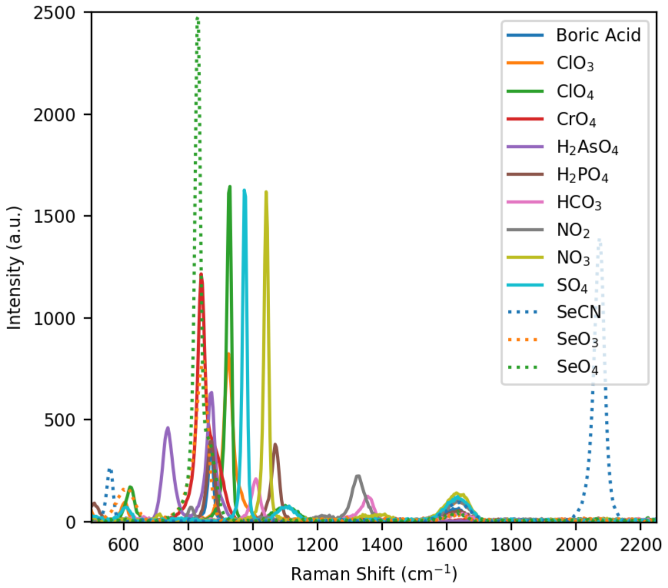


Figure 1: Raman spectra of aqueous anions common to FGD wastewaters.

Figure 2 shows Raman spectra of select species in a range of concentrations, with every 100 measurements averaged together. These spectra (along with others) were used as the basis of ML training datasets to allow for classification of chemical constituents and quantification down to ppm (mg/L) levels. Lower limits of detection (LOD) for each species are provided in Table 2.

Table 2: Example lower limit of detection based on 99% confidence interval from ML, for the utilized training dataset

Chemical Component	LOD (mg/L)
Selenate as Se(VI)	2
Selenite as Se(IV)	4
Selenocyanate as Se(-II)	4*
Nitrate as N	14
Nitrite as N	28
Chlorate	35
Perchlorate as Cl	18
Sulfate	48
Orthophosphate as P	33
Bicarbonate	122
Boric Acid as B	11
Arsenate as As(V)	3*
Chromate as Cr(VI)	.041*

*Detection limits of these species have not yet been fully investigated and are thus likely lower.

When using ML algorithms, a common step in feature extraction is to separate the data into categories or numerical values of interest. PCA was applied to the training data above to reduce the number of features that the model had to process. This improved the efficiency of the model as it did not have to process as many pixels. PCA is a commonly used technique which transforms the features to new coordinate systems optimized to explain as much of the variance as possible among the different features of the data. By finding the features most effective in explaining the data, the method increases interpretability while at the same time minimizing information loss.

After such processing, the data is reduced to a subset of components that are ranked by those that best explain variance. Figure 3 shows only three principle components which explained more than 99% of the variance, and the different chemical target groupings (by color) can be more easily discerned. By such a method, using more transform components, both the type and concentration of a target can be determined using the subsequent classifier algorithms applied to the transformed data.

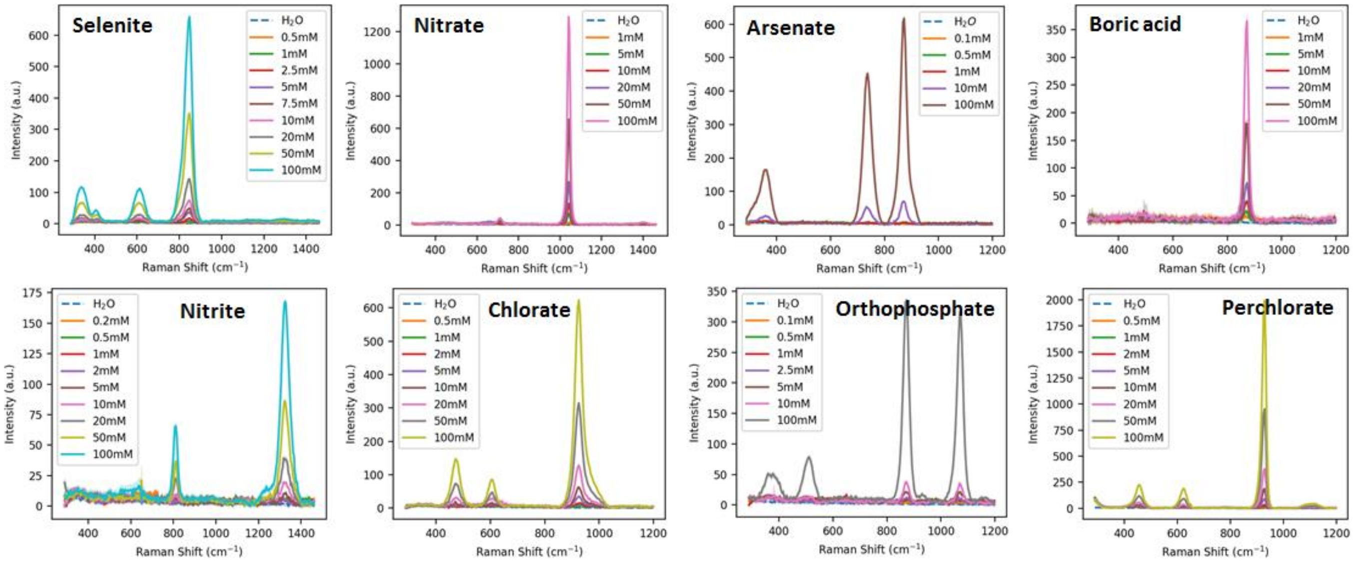


Figure 2: Selection of Raman spectra gathered as ML training data for classification and quantification.

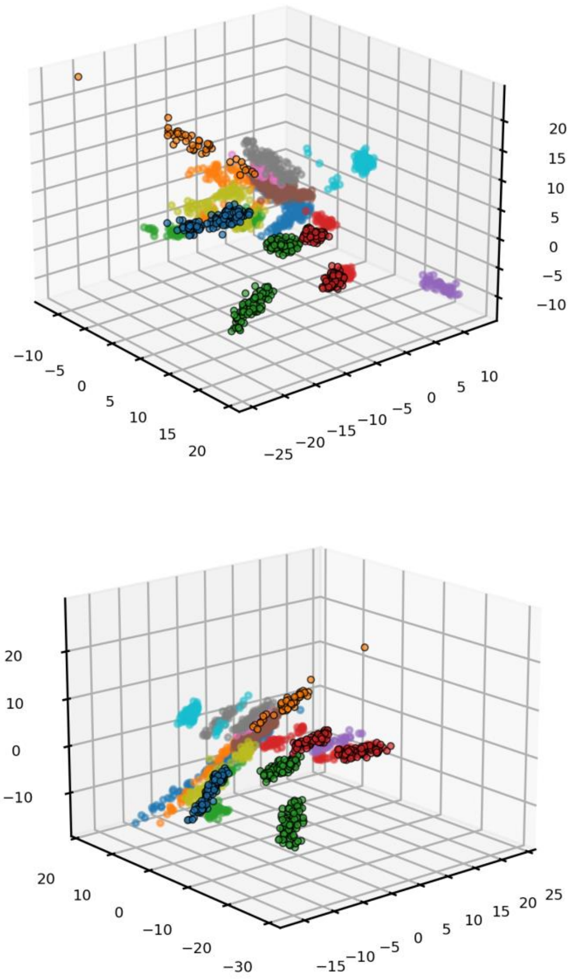
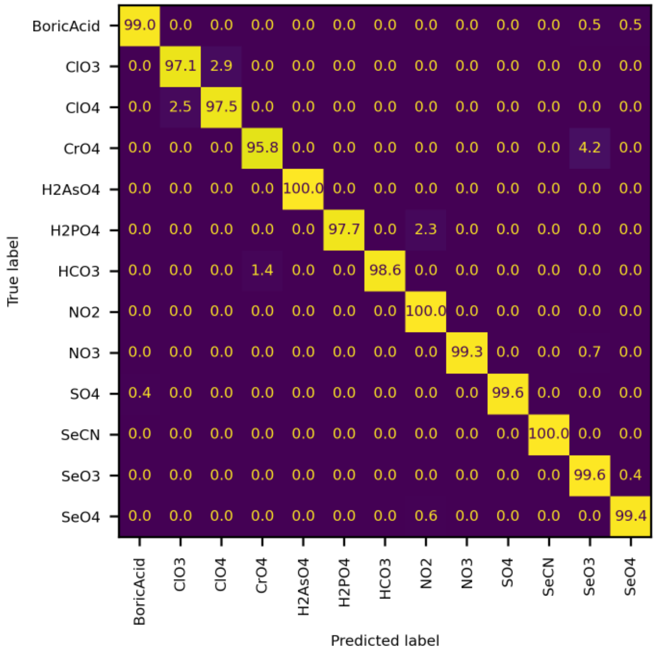


Figure 3: First three features of the data after dimensionality reduction, presented from two different perspectives to show grouping separation.

Figure 4 shows the classification results in the form of a Confusion Matrix [6] for the thirteen species analyzed in the PCA preprocessing, when test data was passed through a SVC model. The Confusion Matrix is a table layout to help visualize the performance of the classification model. Each row of the matrix represents the instances in an actual class, while each column represents the instances in a predicted class. The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another). Matching rows and columns indicate the model's predictions lined up with the truth. Where the prediction matches with a different row or label, then that incorrect classification is also evident on the Confusion Matrix.

For this classifier, we observe greater than 95.8% accuracy on all chemical species. For example, a class with 97.5% accuracy, such as chromate, may be confused with perchlorate 2.5% of the time. This is just one example training-test dataset, and this accuracy is improved with more training data.



	BoricAcid	ClO3	ClO4	CrO4	H2AsO4	H2PO4	HCO3	NO2	NO3	SO4	SeCN	SeO3	SeO4
BoricAcid	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.5
ClO3	0.0	97.1	2.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ClO4	0.0	2.5	97.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CrO4	0.0	0.0	0.0	95.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.2	0.0
H2AsO4	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
H2PO4	0.0	0.0	0.0	0.0	0.0	97.7	0.0	2.3	0.0	0.0	0.0	0.0	0.0
HCO3	0.0	0.0	0.0	1.4	0.0	0.0	98.6	0.0	0.0	0.0	0.0	0.0	0.0
NO2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
NO3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.3	0.0	0.0	0.7	0.0
SO4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.6	0.0	0.0	0.0
SeCN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0
SeO3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.6	0.4
SeO4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	99.4

Figure 4: Confusion matrix showing classification results of the sample dataset of 13 species in water. A value of 100 along the diagonal cells indicates 100% matching of the predicted label with the true label. Values are rounded for visual clarity.

Figure 5 shows an example of classification of a true mixture of FGD wastewater components. In this case, a Raman spectrum was measured of FGD wastewater effluent prior to physical/chemical treatment, provided by the Electric Power Research Institute and taken from a fossil fuel plant. The Sporian system was able to successfully classify three of the prominent Raman-active wastewater constituents. Further training data is being generated to continue to characterize and quantify complex mixtures of wastewater constituents.

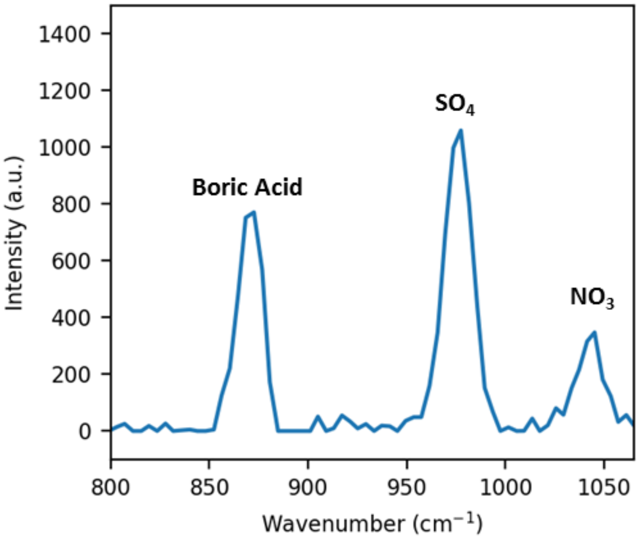


Figure 5: Raman spectrum of filtered, early-stage FGD wastewater effluent.

Conclusions

Contaminated wastewaters are a challenging environment for chemical composition monitoring. Sporian Microsystems' Raman spectroscopy-based measurement system is an effective tool to provide in-line, real-time monitoring of wastewater compositions, and provide constituent classification and quantification. Instruments capable of continuous monitoring in harsh conditions and allowing for the use of ML-based data processing are ideally suited for performing such measurements.

References

1 Water and Wastewater Treatment Market by Offering (Treatment Technologies {Membrane Separation, Membrane Bio-Reactor}, Delivery Equipment, Treatment Chemicals, Instrumentation), and Application (Municipal, Industrial) - Global Forecast to 2027. June 18, 2020, Source: Meticulous Market Research Ltd.

2 Azimi, A., Azari, A., Rezakazemi, M. and Ansarpour, M. "Removal of Heavy Metals from Industrial Wastewaters: A Review." ChemBioEng Reviews, 4 (2017): 37-59.

3 Industrial Wastewater Treatment. International Water Association Publishing (2021, September 28). Retrieved from https://www.iwapublishing.com/news/industrial-wastewater-treatment.

4 Michalski, R. "Ion Chromatography Application in Wastewater Analysis." Separations, 5 (2018).

5 Pedregosa, F. E. A. "Scikit-learn: Machine Learning in Python," Journal of Machine Learning research (2011): 2825-2830.

6 Wikipedia contributors. "Confusion matrix." Wikipedia, The Free Encyclopedia, 23 Sep. 2019. Web. 17 Oct. 2019.