Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2019-377-AC2, 2020

Atmospheric Measurement Techniques Discussions

Interactive comment on "Quantification of toxic metallic elements using machine learning techniques and spark emission spectroscopy" *by* Seyyed Ali Davari and Anthony S. Wexler

Seyyed Ali Davari and Anthony S. Wexler

sadavari@ucdavis.edu

Received and published: 17 February 2020

A new approach is presented to detect toxic metallic elements in the atmosphere. The approach is based on spark emission spectroscopy. The authors develop a new spectrometer that they claim is more cost-effective than previous detectors. They record spectra for Cr, Cu, Ni and Pb at different concentrations and then deploy the Least Absolute Shrinkage and Selection Operator (LASSO) machine learning method. It is still not clear to me why they apply LASSO, presumably to calibrate their spectrometer. The approach seems interesting and promising. However, the manuscript is quite unstructured and I had to deduce the main objective since it is not clearly stated. The

manuscript is quite immature and in its current state not suitable for publication. There is no flow and logical connection between sections and results are reported out of place and in the wrong order. Maybe some of the statements are clear to members of the atmospheric community, but they were not clear to me.

Authors' Response: The authors would like to thank the reviewer for his/her time during the peer-review process. We also would like to state that we have not developed a detector. The paper presents our investigation for development of a spark emission spectroscopy system that uses Ocean Optics spectrometer as the system detector. Moreover, LASSO is a regression technique that has been used for data analysis, not detector calibration.

2. Table 1 lists hazardous elements with their full names, whereas Table 2 only gives the chemical elements. For consistency this should be changed. Also, chromium, nickel and lead show up in both tables, which leaves only copper as unique element in Table 2. This is strange.

Authors' Response: The authors thank the reviewer for pointing out the error. In the revised manuscript we have addressed the reviewer comment.

3. "Table 2 lists other metals that are not on US EPA's HAPs list but have been implicated in a range of adverse health effects so are of concern to the California Air Resources Board (CARB). X-ray fluorescence (XRF) and inductively coupled plasma mass spectrometry (ICP-MS) have been used traditionally to quantify metals in atmospheric particles." This is just one example of the structural and systemic problems of this article. From a list of hazardous elements the authors jump straight, in the same paragraph, to detection methods for such elements without establishing first that an important task or objective is to measure metal elements in the atmosphere. Maybe I am being picky here, but I had difficulties reading the introduction and following the logic.

Authors' Response: The authors would thank the reviewer for the directions to improve

the manuscript. In the revised manuscript, we have addressed the issue to improve the introduction readability.

-"LOD": the abbreviation LOD in the introduction is not explained.

Authors' Response: The authors have added the definition of LOD in Pg. 2, Line 37: "... is expensive, has a high limit of detection (LOD) for heavier elements,..."

The introduction essentially consists of an endless list of previous studies. At the end the reader is none the wiser, because no assessment or reflection is given. Worse; after this endless and tedious list the authors say "In this study we employed spark emission spectroscopy to quantify toxic metallic elements." At this point the readers wonders "so what?". Another methods for detection. What is new?

Authors' Response: The main goal of this study is to develop a low-cost spark emission system and improve the analytical performance using advanced data analysis techniques such as K-Means clustering and machine learning. In order to address the reviewer concern, we have added the following at Pg. 3, Line 59: 'While LIBS and SIBS address issues regarding the field measurement...'

The introduction does not mention LASSO or machine learning at all. Since this is quite a large part of the paper, it should be mentioned. There are plenty of good overview articles for machine learning to refer to, for example. - When asked to review the manuscript, my interest was piqued by the prospect of machine learning for spectra. Most machine learning is done for single target properties and machine learning for spectra is quite difficult. Examples are: J. Timoshenko, A. Anspoks, A. Cintins, A. Kuzmin, J. Purans, A. I. Frenkel, Phys. Rev. Lett. 2018, 120, 225502. C. Zheng, K. Mathew, C. Chen, Y. Chen, H. Tang, A. Dozier, J. J. Kas, F. D. Vila, J. J. Rehr, L. F. J. Piper, K. A. Persson, S. P. Ong, npj Comput. Mater. 2018, 4, 12. K. Ghosh, A. Stuke, M. Todorovic, P. B. Jorgensen, M. N. Schmidt, A. Vehtari and P. Rinke, Adv. Sci. 6, 1801367 (2019) A. Cui et al Phys. Rev. Applied 12, 054049 (2019) W. Lee, A. T. M. Lenferink, C. Otto, H. L. Offerhaus, J. Raman. Spectrosc. 1 (2019) But then I saw no

discussion of machine learning for spectra in this article.

Authors' Response: The authors would like state that the main focus of the current research was development of a low-cost spark emission spectroscopy to detect and quantify toxic metal PMs in atmosphere. Compared to the pervious studies, the expensive components such as spark generation and delay generator have been developed by the authors. The low-cost components such as delay generator might show false readings in some instances. We employed advanced machine learning techniques such as K-Means clustering to detect those false readings and discard them in order to clean the spectroscopic dataset and consequently reduce the errors. Moreover, most of metallic transitions occur in UV-VIS region of the spectrum. Using a low-cost spectrometer, we would not be able to resolve the spectrum sufficiently to detect individual metallic peaks to use them for quantification. Therefore, it becomes challenging to identify features in the spectrum that might be used for quantification. Instead, we chose LASSO as our data analysis technique. LASSO has the advantage that is not limited to individual peaks and performs the feature selection automatically and hence more suitable for identify metallic elements. In order to address the reviewer comment, we have added the following in Pg. 3, Line 59: "While LIBS and SIBS address issues regarding..."

Instrument development: this part appeared strange and out of place to me on my first reading of the manuscript. At no point had I been prepared for a long, technical description of a new spectrometer. I think this is mostly a flow and logic problem again that can be solved by having a few connecting sentences that guide the reader through the paper.

Instrument development: this part appeared strange and out of place to me on my first reading of the manuscript. At no point had I been prepared for a long, technical description of a new spectrometer. I think this is mostly a flow and logic problem again that can be solved by having a few connecting sentences that guide the reader through the paper.

Authors' Response: As it was mentioned, the main goal of this study is to develop a low-cost SIBS system and hence the authors provided the details related for instrument development. In order to address the flow of the manuscript, the authors have added the following in Pg. 3, Line 59: "While LIBS and SIBS address issues regarding. . ."

- Figure 4 shows the expected delay as a function of the measured delay. I don't quite understand what that tells me or why that is important, but first of all, what are the circles in Figure 4 and what is the red dashed line. Second, how does one get from this expected or measured delay to a spectrum as shown in Figure 9?

Authors' Response: In time resolved spectroscopy, usually a delay generator is needed to resolve the spectrum temporarily. We have designed and developed a delay generator to reduce the cost. Figure 4 shows the performance of our delay generator. The Y axis illustrates the delay values that we set with the delay generator, and X axis shows the delay values that we measured using an oscilloscope. The circles indicate the measured values with the oscilloscope and the red dash line indicates the one-to-one ratio line. To clarify the Figure, we have added the following at Pg. 4, Line 103: "Fig. 4 shows the delay generator performance. The Y axis illustrates the delay values requested of the delay generator while the X axis shows the measured values. The red dashed line shows the desired 1:1 line while the circles show the measured performance. The performance is linear over with a slight deviation from the 1:1 line."

"an unsupervised learning technique, K-means clustering. . ." how is the K-means applied? No details are given.

Authors' Response: The K-Means clustering performed in order to discard the "outliers" from the spectra dataset. As it was stated in the manuscript, the first step is to determine the number of clusters. This has been performed by plotting the within-cluster sum of squares (WCSS) as a function of number of clusters. Obviously, increasing the number of clusters will reduce the error. However, this might lead to overfitting problem. It is standard to set the optimum number of clusters to the value, where WCSS error

becomes plateau. This has been shown in Figure 5. In order to address the reviewer concern, the following has been added in the revised manuscript at Pg. 5, Line 120: "The standard approach is to set the optimum number of clusters to the value where the within-cluster sum of squares (WCSS) error plateaus."

"The standard approach is to set the optimum number of clusters to the value where the within-cluster sum of squares (WCSS) error plateaus."

Authors' Response: The spark emission spectroscopy is based on ablating materials using high voltage-current system. Since the voltage and current are very high, they create electromagnetic interference that might affect the delay generator and other electronic components. This results in noise in the electroncis and hence generates outliers in the dataset. To address this issue, we employed K-Means clustering to identify outliers in the dataset. For example, the following graph shows the spectrum of Cr obtained after 2μs delay: However, if electromagnetic fields generated by the spark interferes with the electronics altering the delay value, the following spectrum results: as it can be observed the second spectrum is completely different from the normal spectrum and incorporating the second in our analysis will add error to the further analysis. K-Means clustering ensures us that the cleaned dataset will not contain erroneous spectra thus improving the accuracy and precision of the linear models.

The results section then jumps straight to LASSO without saying why LASSO is applied. It is still not 100% clear to me. What is actually measured by the spark emission spectrometer and why does one need machine learning?

Authors' Response: The goal of the study is to detect and quantify toxic metal concentration in atmosphere. The spark generates plasma that excites toxic metals. Once they relax back to ground states they emit the orbitals energy difference as light. A fiber optics collects the light and transmits it to a spectrometer that resolves the light into different wavelengths. We use the resulted spectrum to detect and quantify the concentration of toxic metal. In order to quantify the concentration of these pollutants,

we need to have a model that receives an input (i.e., a peak at a specific wavelength, multiple peaks, entire spectrum) and maps the input to the concentrations. This mapping can be generated using linear regression, neural networks, Gaussian process, etc. We chose LASSO as the model that receives the entire spectrum and maps it to concentration of the pollutants. LASSO compared to other techniques such as partial least square (PLS) can determine, which features (wavelengths) are more correlated to the output. This means that it only keeps a few features and discards the rest of features. In this study, our Ocean Optics spectrometer has 2048 pixels, which means that the recorded spectrum has 2048 features. Let's denote the entire spectrum as $\mathbf{x} \in^{2048}$. We can consider the entire spectrum as a high-dimensional vector. Our goal is to develop a mapping between this highly dimensional vector and concentration values:

$$h : \mathbf{x} \in^{2048} \rightarrow C \in \tag{1}$$

LASSO compared to other regression models only uses a few features of the high-dimension vector to generate the linear model. It is worth mentioning that one of the main reasons to use ML was the spectrometer poor resolution. The current spectrometer does not have sufficient resolution to resolve close peaks. As a result, the peaks can convolute to each other and hence it is impossible to develop a model based on known emission peaks. In order to address the reviewer concern, the following has been added to the revised manuscript at Pg. 6, Line 139: "The cleaned scaled spectra set has been used to detect and quantify concentrations of the toxic metals."

Figure 6 has two insets that are way too small to be readable. Moreover, I do not understand what I see in the Figure and the caption is confusing. The points (how have they been determined) seem to cluster in a red region and a purple region that is barely visible. What does this mean and what is then done with that information?

Authors' Response: Figure 6 illustrates the effectiveness of K-Means clustering in detecting outlier spectra. As it was explained in the previous questions, K-Means clustering has been used to identify outlier spectra and exclude them from the LASSO. As

it was explained, each spectrum can be regarded as a high-dimensional vector, which each component of the vector indicates the intensity at a specific wavelength. The outcome of K-Means will be normal spectra set that has excluded the outliers from the spectra set.

Equation 1 and its explanation make no sense to me. What is x? The discretized x-axis, in other words the wavelength values? h theta is apparently the normalized spectrum, but why does it depend on the LASSO coefficients theta? y(i) is the known concentration corresponding to spectrum i, but in equation 1 h theta is subtracted from y(i). How can a spectrum be subtracted from a concentration?

Authors' Response: The followings summarize the terms:

- $\mathbf{x}$ indicates the intensities correspond to each wavelength.
- $h : \mathbf{x} \in^{2048} \rightarrow C \in$ The function with θ parameters (to be determined) that maps spectrum intensities to concentration.
- $y^{(i)}$: The concentration corresponds to spectrum i_{th}.

To address the reviewer comment, the following has been added to the revised manuscript at Pg. 7, Line 140: '... $\mathbf{x}^{(i)} \in^{2048}$ and $h_\theta(\mathbf{x}^{(i)})$ represent...'

"Therefore, we set the regularization constant to the value that minimizes the loss for the test set." And what is that value? It should be reported. C is a hyper parameter and hyper parameters are an essential part of machine learning.

Authors' Response: The authors would like to thank the reviewer for pointing out the issue. In the revised manuscript the hyper parameters for various elements has been reported in Table 3.

Figure 9 now suddenly shows a spectrum. More like an afterthought. How would one actually extract the mass from such a spectrum?

Authors' Response: The mass is predicted based on the model that receives the spectrum as an input. Figure 10 illustrates the features that have been selected by LASSO model (red lines) and compares it with the original spectrum. The goal was to show how LASSO effectively chose less number of features and used them for developing a predictive model.

Machine learning features heavily in the manuscript. However, at no point do the authors demonstrate that their method actually learns, i.e. its accuracy improves with more data. It is now standard to show learning curves in machine learning work. A learning curve plots the target (e.g. the prediction accuracy) as a function of training data size. The predictive accuracy should increase with increasing training set size (i.e. the error in the prediction decreases)

Authors' Response: We actually have demonstrated the learning process in Figure 8. Figure 8 shows the loss value as a function of number of features. It is well known that as the number of features increases, the model over fit the data. The loss values for the training set indicate this phenomenon perfectly. Moreover, considering the loss values for the test set, we realize that the error increases after incorporating certain number of features, which suggests the optimum number of features.