Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2019-377-RC2, 2019

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

Anonymous Referee #2

Received and published: 19 December 2019

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.

More detailed comments are given below.

- 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.

- "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.

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

- 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?

- 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, D. Lu, Y. Lin, A. I. Frenkel, J. Phys. Chem. Lett. 2017, 8, 5091.

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. Todorović, P. B. Jørgensen, 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.

- 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.

- 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?

- "an unsupervised learning technique, K-means clustering..." how is the K-means

applied? No details are given.

- It is not clear to me why a clustering technique is applied in the first place. The authors say this is to remove outliers. But what outliers? How do they manifest and why are they there in the first place? Should one remove them even?

- 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?

- 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?

- Equation 1 and its explanation make no sense to me. What is x? The discretised x-axis, in other words the wavelength values? h_theta is apparently the normalised 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?

- "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.

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

- 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).

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2019-377, 2019.