Modeling Respirable Silica

A Collaboration between CS96 and CDC's NIOSH

Ashley Zhuang, Eric Lin, Julia Curran, & Troy Appel

Harvard Faculty of Arts and Sciences

Problem and Goals

For miners and construction workers, a major health concern is inhaling a compound called respirable crystalline silica (RCS), which is found in dust in their working environments. When inhaled, RCS has been shown to cause permanent and potentially fatal lung diseases, such as silicosis. Currently, the process of detecting RCS levels consists of collecting dust or air samples at the work site and then sending this data to a laboratory for analysis — a process that can take days to weeks. During this time, workers may be inhaling dangerously high levels of toxins that just have not yet been detected. Consequently, our work aims to develop a model to estimate RCS levels in real-time. We focus our efforts in estimating quartz levels, as quartz mineral deposits are one of the most common naturally occurring forms of silica found in coal mines.

Data

Data Collection

The main dataset contains the Fourier transform infrared (FTIR) spectra for 890 respirable dust samples, synthetically created at the NIOSH Pittsburgh Mining Research Division (PMRD) facilities. The first 582 samples have only quartz present, while the other 382 samples have other minerals introduced as confounders. We were also given the weight of the quartz along with the weight of the confounding mineral (if applicable) in this dataset. These 13 mineral mixtures were used as confounders: albite, amorphous silica, anorthite, chlorite, cristobalite, dolomite, kaolinite, magnetite, muscovite, oligoclase, pyrite, talc, and tridymite.

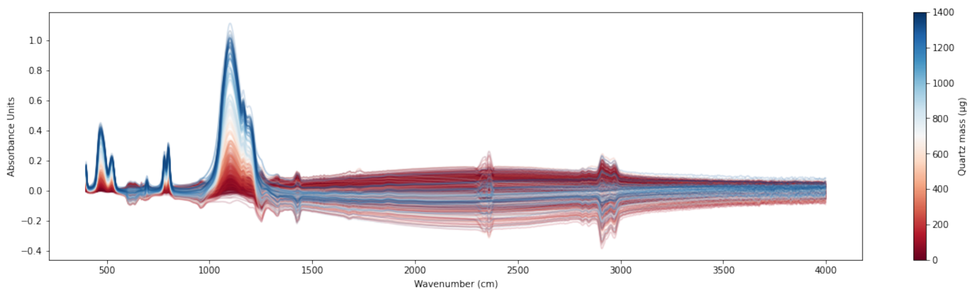

Figure 1: Spectral data of quartz-only samples.

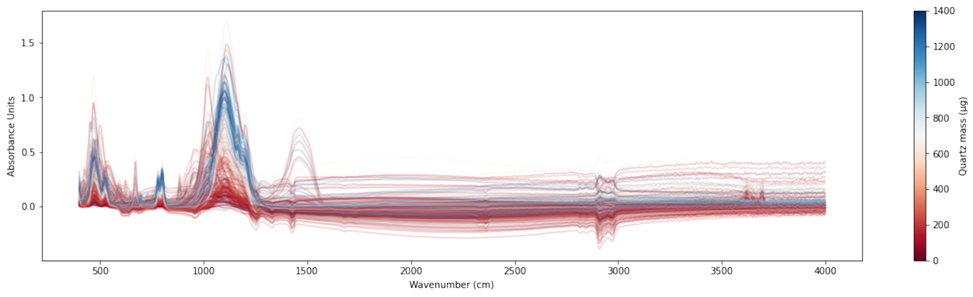

Figure 2: Spectral data of quartz-plus-confounder samples.

Data Preprocessing

Sanitizing data required two mechanisms. First, we needed to remove variance caused by filter miscalibration. This was done by assuming a spectrum resulted from an average filter, and finding the 'best' deviation from an average filter to explain a spectrum. This 'best' deviation was the one that minimized the spectrum's second derivative around spectroscopic peaks of the filter, which we found to be a good proxy for spectral complexity.

After removing the influence of the filter, FTIR is still liable to introduce some background spectrum to our empirical data. There are several algorithms from the literature which we adapted and modified to suit our needs. Among these were:

(i) Asymmetric Least Squares [2]

(ii) Iterated Discrete Wavelet Transform (DWT) removal [4]

(iii) Iterated Polynomial removal

Of these, it was found that the polynomial removal was the only method that did not change the characteristic shape of the spectra in a substantial way.

Data was additionally smoothed via a Gaussian blur to reduce sensor noise. Similar accuracy was found across a wide range of blur amounts.

Feature Engineering

Using absorbance alone is often not sufficient to distinguish components. The following is a list of tested data transformations and characterizations of each.

(i) First and second numerical derivatives with respect to wavenumber. These have the effect of removing constant and linear terms, reducing the effect of the background spectrum. [3]

(ii) Discrete wavelet transform (DWT). This provides multi-resolution data local to every wavenumber, sufficient to reconstruct the original data entirely.

(iii) Continuous wavelet transform (CWT). This provides multi-resolution data local to every wavenumber, and allows a wider variety of wavelets to be used, plus more desirable smoothness properties.

Methods and Models

Regressive Models: Partial Least Squares & Ridge Regression

Our baseline model is Partial Least Squares, which is employed by previous research literature, as a simple regression model. PLS works by building a linear regressor that fits projections of predicted and observed variables to a new (lower dimensional) space. Due to this projection, PLS is suited to working with datasets where the number of features is greater than the number of observations.

We also employed Ridge Regression, which is a common method used in chemistry and econometrics where variables may be highly correlated. It accomplishes this through introducing a regularization with a penalty term associated with the sum of squared errors.

Generative Model

All spectra are produced via material processes. The shape of spectra can be characterized as the sum of a number of Voigt functions, which are convolutions of Gaussian and Cauchy curves [1]. Using an iterated process, we are able to find the minimal confounding spectra to explain a spectrum using a general function optimizer.

This model is not included in results as its use case is not comparable to the given models.

Evaluation

To evaluate our approaches, we used a train-validation-test split of 70:15:15. The validation set was used as the target of experimentation, and we report the test performance only after finishing hyperparameter tuning.

Results

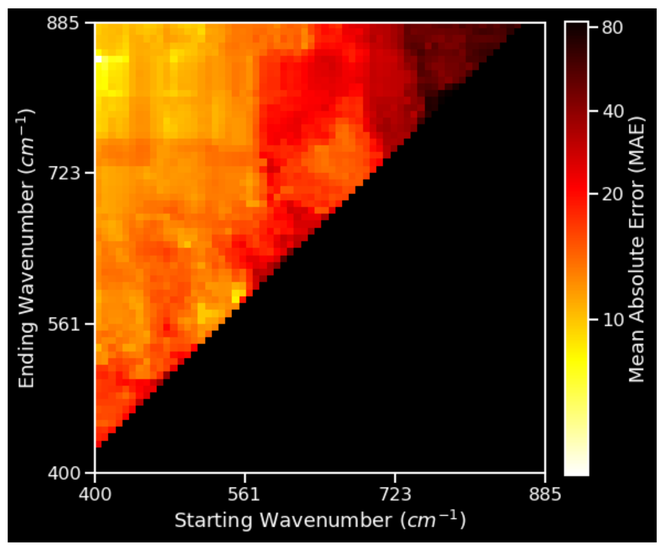

Figure 3: Error of models trained on various subsets of the data. Final model was most accurate when allowed to train on a (nearly) full spectrum.

Chosen models

After iterating over every combination of preprocessor, preprocessor parameters, and model, we concluded that the following preprocessor specification was ideal on the validation set:

(i) Output: CWT, scales [8,16,32], gaussian wavelet

(ii) Fit filter: yes

(iii) Background fit: iterated polynomial degree 5

(iv) Blur strength: 5

(v) Window: wavenumbers 400-860

We then selected the best parameters for PLS and Ridge regression.

Model	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)
Baseline PLS	19.68	16.47
Tuned PLS	17.67	14.53
Default Ridge	20.8	17.3
Tuned Ridge	**16.6**	**14.2**

Table 1: Test results, before and after tuning on the validation set.

Conclusions

In this work, we present improvements for modeling respirable silica:

- Our model pipeline addresses real-world challenges of training on insufficient and unrepresentative data and generalizing to unseen data polluted with confounders.
- We achieve 6.4% better RMSE than the original baseline PLS approach when testing on samples with confounders. This is due to not only model tuning but also important wavelet feature engineering.

Next Steps

We are handing off our work to the CDC for real-world deployment. Anticipating changes in the data distribution from mine samples, we have prepackaged our proposed models into a singular tunable pipeline. Future work can tune parameters such as the specific wavenumber windows on new training data that is closer in distribution to real world testing.

References

[1] B.H. Armstrong. Spectrum line profiles: The voigt function. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 7(1):61–88, 1967.

[2] Sung-June Baek, Aaron Park, Young-Jin Ahn, and Jaebum Choo. Baseline correction using asymmetrically reweighted penalized least squares smoothing. *Analyst*, 140, 09 2014.

[3] J. Dubrovkin. *Derivative Spectroscopy*. Cambridge Scholars Publishing, 2020.

[4] C Galloway, Eric Le Ru, and P Etchegoin. An iterative algorithm for background removal in spectroscopy by wavelet transforms. *Applied spectroscopy*, 63:1370–6, 12 2009.

Acknowledgements

We owe a debt of gratitude to Cody Wolfe, who has graciously supported this project on behalf of the CDC. We also extend special thanks to our Teaching Fellow (TF) Martin Reyes, Professor Jim Waldo, Head TF April Chen, and TF Catherine Yeo for their mentorship.