Applying Machine Learning to Determine the Biological Source of Atmospheric Aerosols

Nathaniel Breaux, Gregory Vandergrift, Ashfiqur Rahman, Swarup China
Pacific Northwest National Laboratory - Environmental Molecular Sciences Laboratory - Richland, Washington, USA
Contact: nabreaux42@gmail.com

Motivation

- EMSL's MONet project will require quick classification of atmospheric samples
- Climate sciences depends on accurate aerosol compositions for radiative balances
- Many aerosols seen in field campaigns have unidentified sources or have not been produced in chamber studies
- Machine learning has been used in other fields to classify aerosol types[1]

Atmospheric Samples

- Figure 1 displays the source location and times of the samples used for model training and validation
- Samples collected throughout the year to display both isoprene- and terpene-dominated environments

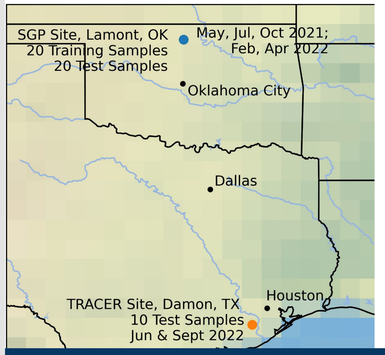


Figure 1. The site, collection period, and number of samples included in the SGP and TRACER sets. Samples were collected at day and night at both sites. The training set trained the models, and the testing sets were used to evaluate models by comparing predictions to true labels.

Data Preprocessing

- Atmospheric samples previously processed using mass spectrometry
- Most likely chemical formula assigned using MFAssignR
- Only CHOS (Organosulfate) group analyzed
- Data manually assigned to groups using Surratt paper on chamber experiments[2]

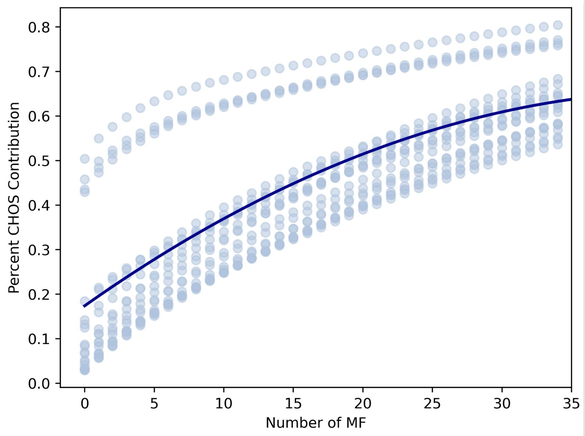


Figure 2. Percent contribution of x-most abundant organosulfates to the total abundance. Used to determine the number of aerosols to label in each file for training.

Models

- Used Python scikit-learn library to train decision tree, random forest, gradient boost, and MLP classifier
- Default parameters used to initially train models on whether aerosols are isoprene-related, terpene-related, or unknown
- Figure 2 shows a schematic of the model prediction process
- Evaluation metrics such as accuracy and F1 score used to determine best performing model
- The top model's hyperparameters were tuned using a grid search algorithm

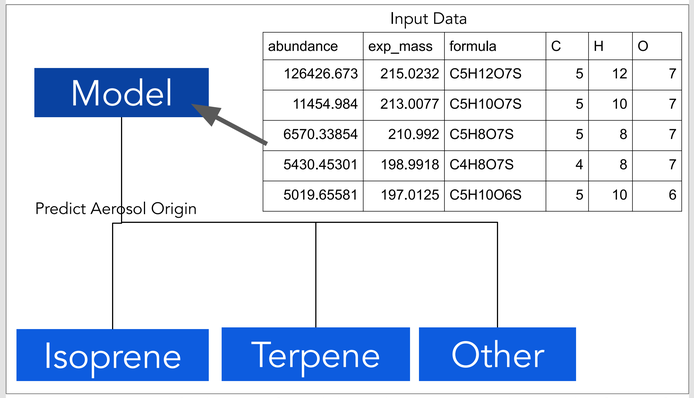


Figure 3. A simplified view of model prediction. Input data in the form of specific quantities and chemical qualities of aerosols is used to predict if each aerosol is isoprene-related, terpene-related, or unknown (other).

Model Results

Table 1. Model prediction evaluations for the 20 SGP test samples

	Decision Tree	Random Forest	Gradient Boost	MLP Classifier
Accuracy	0.96	0.97	0.96	0.52
F1 Scores by Class	[0.98, 0.92, 0.97]	[0.98, 0.92, 0.97]	[0.95, 0.94, 0.97]	[0.56, 0.16, 0.63]

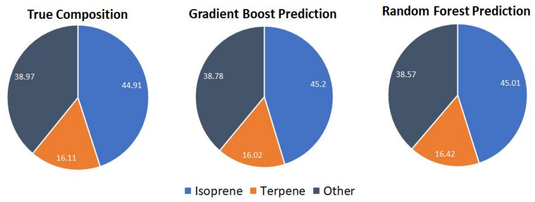


Figure 4. Percent compositions for the two best-performing models compared to the true composition of the 20 SGP test samples

Table 2. Model prediction evaluations for the 10 TRACER samples

	Decision Tree	Random Forest	Gradient Boost	MLP Classifier
Accuracy	0.93	0.94	0.94	0.74
F1 Scores by Class	[0.95, 0.88, 0.94]	[0.96, 0.90, 0.92]	[0.93, 0.92, 0.95]	[0.59, 0.77, 0.79]

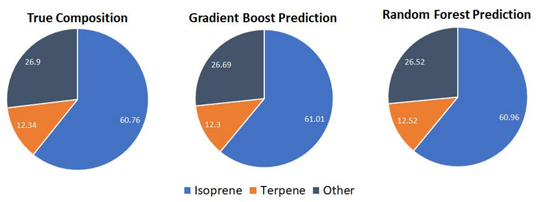


Figure 5. Percent compositions for the two best-performing models compared to the true composition of the 10 TRACER samples

Unidentified Aerosol Predictions

- Some previously unidentified aerosols were predicted by the gradient boost model as isoprene- or terpene-related
- Used linear regression to find correlations between the abundance of these aerosols and aerosols identified in the Surratt paper
- R^2 values greater than 0.7 regarded as possibly related and should be examined further
- Heatmaps below show R^2 values for linear regression between unknown and known aerosols

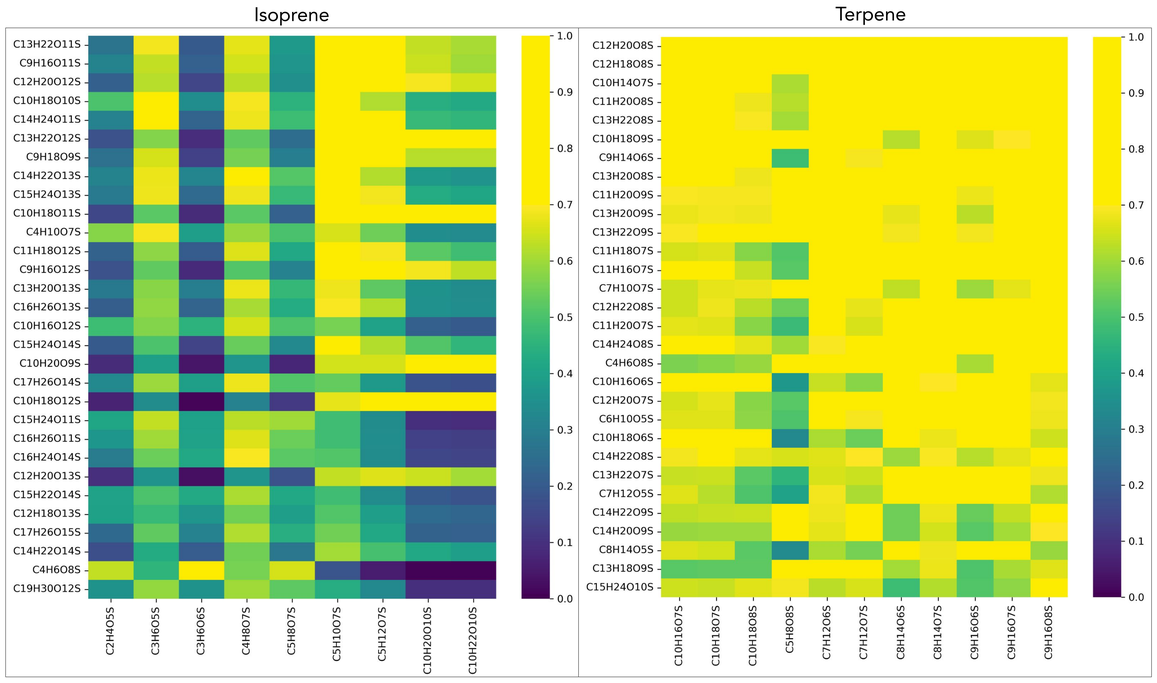


Figure 6. Heatmaps showing R^2 values for isoprene (left) and terpene (right). Unidentified aerosols are on the vertical axis, and known aerosols are on the horizontal axis. R^2 values greater than 0.7 are uniformly yellow as they should be investigated further. Different known aerosols tend to occur in different environments/times of year, so correlation may not exist uniformly across all known aerosols.

Conclusions and Future Work

- Machine learning models can be used to identify biogenic source of atmospheric organosulfate aerosols effectively
  - Gradient boost was the best model overall
- Aerosols with strong correlations to known aerosols should be searched for in future atmospheric sampling
- The "other" class should be better defined through further chamber studies
- Long term plans should expand beyond classifying only organosulfates - this will require more chamber studies and atmospheric samples to train models

Acknowledgements

Thank you to Sonja Moons for creating the MFAssignR-processed files used to train and test the models. This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program.

References

[1]Laura Van Hese et. al., "Towards real-time intraoperative tissue interrogation for REIMS-guided glioma surgery", J. Mass Spec. & Adv. Clinical Lab 24, 80-89 (2022)
[2]Jason D. Surratt et. al., "Organosulfate formation in biogenic secondary organic aerosol", J. Phys. Chem. A 112 (36), 8345-8378 (2008)