# Identifying the Influence of Local Source Emissions on the Regional Representativeness of AOS Measurements Using Machine Learning

**Jeffery Mitchell, Scott Smith, Laurie Gregory, Thomas Madigan, Andrew McMahon, Stephen Springston, Richard Wagener**

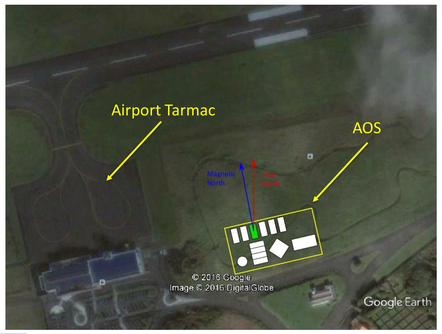
***Brookhaven National Laboratory, Upton, New York***

## Local Source Emissions at the ENA site

The Aerosol Observing System (AOS) at the ENA site in the Azores off of Portugal is located next to an airport (GRW).

The location of the AOS is shown schematically in the figure on the right.

In order to monitor airport activity and its affect on AOS data, a tower camera was installed at the site and aimed at the tarmac.

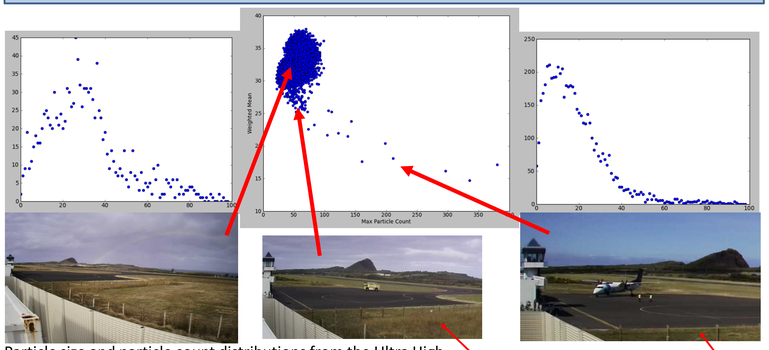



There are typically 2 SATA Air Acores flights to and from the airport each day along with the occasional cargo flight.

A fire truck and several cars perform a safety inspection of the runway every morning.

**A data-based machine learning application has been developed to automatically identify AOS instrument data affected by local source emissions.**

## Local Source Emissions Observed in AOS Data

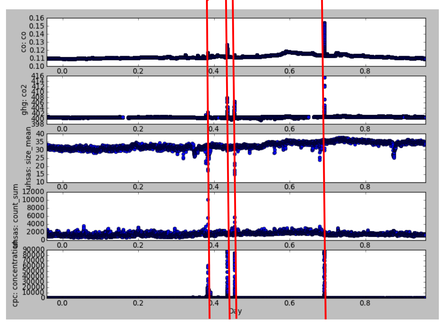



Particle size and particle count distributions from the Ultra High Sensitivity Aerosol Spectrometer (UHSAS) are sensitive to local source emissions. There is some discriminating power between airplane and truck exhaust.

Shown here are time series data from four distinct AOS instruments from a single day. The measurements include:

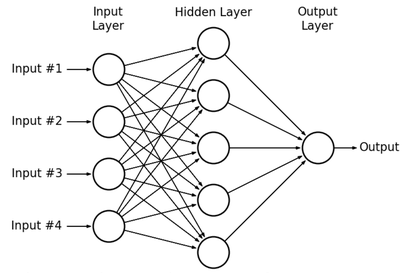
- Carbon monoxide concentration
- CO2 concentration
- Condensation Particle Counter (CPC) aerosol concentration
- UHSAS mean particle size and particle count integral.

The large spikes are correlated with local source emissions from the airport.

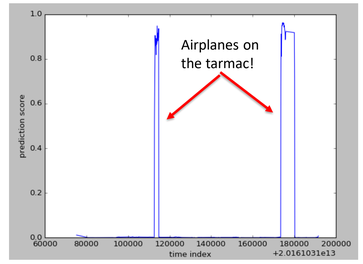



## A Neural Network For Airplane Recognition

- For supervised machine learning, data influenced by local source emissions must be tagged in the training set before training the model.
- To identify times affected by airplane exhaust, a neural network application has been developed to automatically recognize airplanes on the tarmac in images taken by the tower camera.
- The images are cropped, pixelated, and gray-scaled. Each pixel serves as a single input to the neural network.
- The neural network is trained using 200 positive images and 550 negative images. It identifies airplanes in single images with an accuracy of 97%.
- **Once trained, the neural network can process a days worth of 700 images on a laptop in about 10 seconds.**



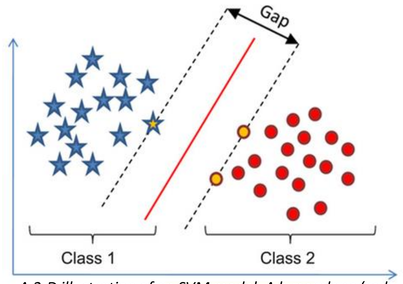

The neural network consists of 2,782 input units, 360 rectified linear hidden units, and a single sigmoid output unit.



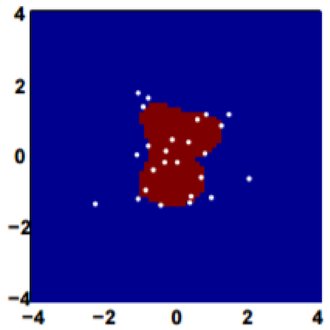

The prediction score from the neural network for 700 images from one day. The two SATA flights are correctly identified.

## A Support Vector Machine Model to Identify Local Source Emissions

- It is most efficient to identify local source emissions using a model capable of incorporating information from multiple measurements from multiple instruments simultaneously.
- A Support Vector Machine (SVM) model is implemented for this purpose.
- SVMs are efficient for well separated classes, can handle non-linear correlations, and can be insensitive to noise in the data.




*A 2-D illustration of an SVM model. A hyperplane (red line) is constructed to maximize the class separation. The yellow points are the support vectors.*



*A 2-D illustration of a one-class SVM model applied to a data set. The red shaded area represents the decision boundary bounded by the constructed hyperplane. White points outside this area are reported as anomalies.*

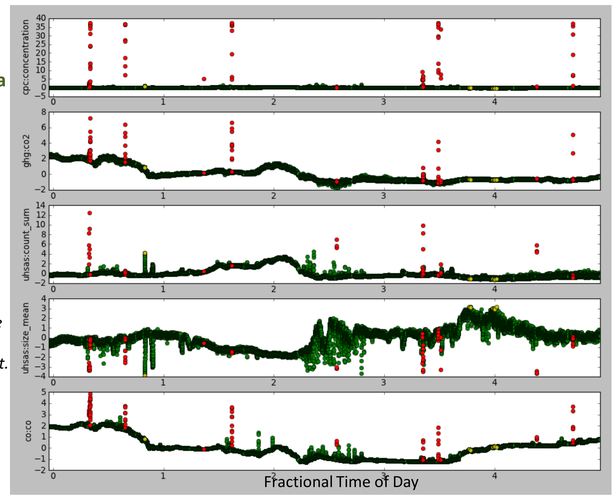
- Since local source emissions in the AOS data influence a small fraction of the total data set, the problem is approached as an anomaly detection problem.
- One-class SVM models are ideal for this approach.
- The model is first trained using data that is free of the influence from local source emissions.
- Then, when processing new data, the one-class SVM provides a binary output of 1 if the data is “normal”, or 0 if it is not (an outlier).

## Training the Support Vector Machine Model

Normal Data

Outliers

False Negatives




*AOS data covering five days. Green points indicate the training set.*

- A training set covering 5 days of AOS operation is defined from 10/26/16-10/30/16.
- All data streams are z-score normalized to optimize the SVM performance.
- The data is synchronized in time to the instrument with the lowest sampling rate. For this case, it is 10 seconds for the UHSAS.
- Data correlated with local source emissions (using the neural network) are removed from the data and the SVM is trained on the remaining data.
- The plot above shows the output of the SVM model after processing the training set.

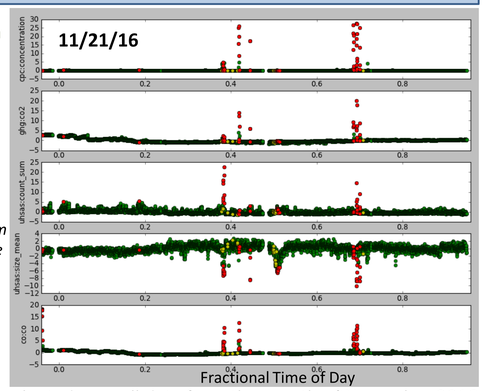
## Evaluating the Support Vector Machine Model

Normal Data

Outliers

False Negatives

*AOS data from 11/21/16. Events from two airplanes and fire trucks are correctly identified as red points.*




- The one-class SVM model was evaluated over all data from the month of November 2016.
- The accuracy for classifying normal data is better than 99%.
- **The accuracy for classifying data from local source emissions is better than 99%.**
- **It takes less than 15 seconds to classify all of the data for one day.**
  Most of that time is spent in I/O and time synchronization.

## Summary

- **An efficient and fast automatic application for the identification of AOS data influenced by local source emissions at the ENA site has been developed.**
- **This application will be able to reduce mentor effort for data quality.**
- Local source emissions are identified using the AOS instrument configuration alone.
- The model incorporates data streams from multiple AOS instruments.
- The application is easily configured for any combination of AOS data streams at any site.