Determining a Relationship Between Two Distinct Atmospheric Datasets of Different Granularities

Sami Benzaid, Janara Christensen, David Musicant, Emma Turetsky

ABSTRACT

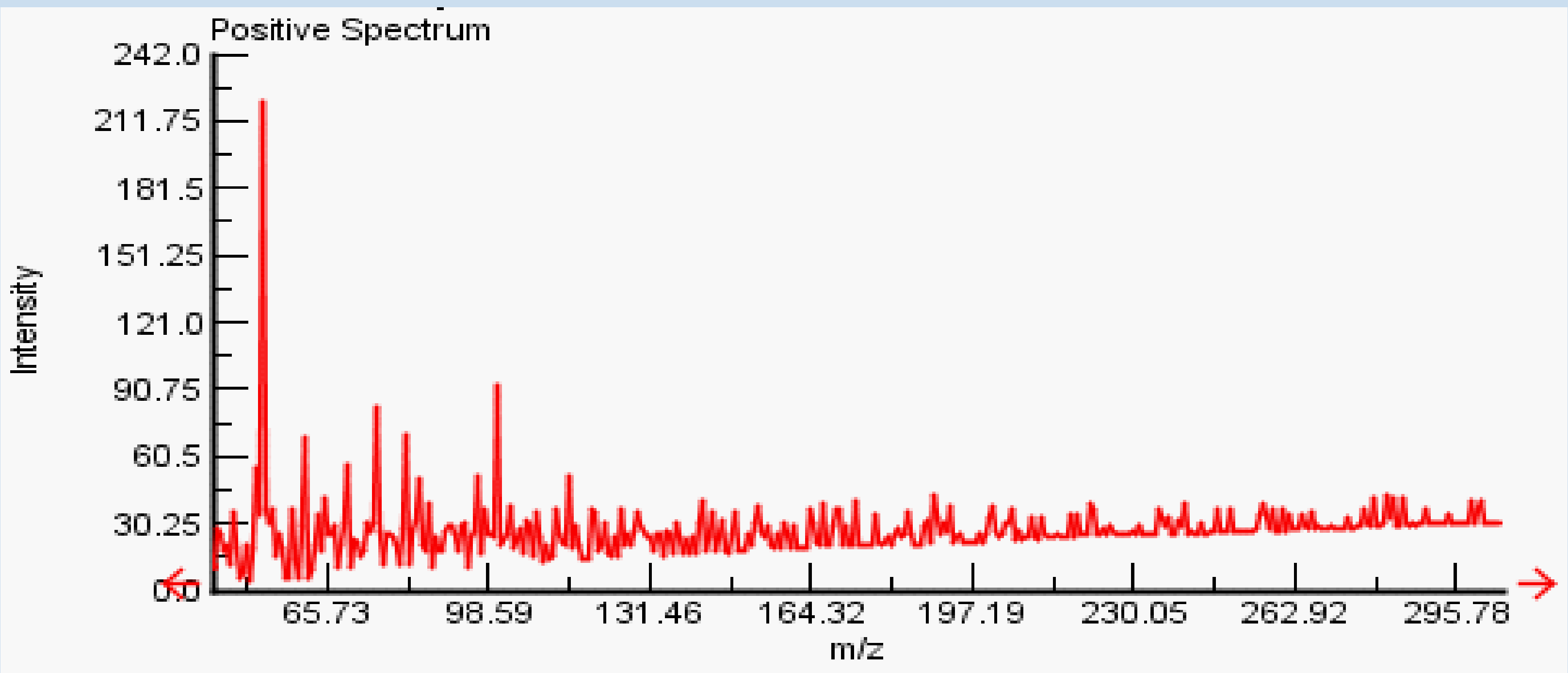
Here, we present several methods of using data mining and statistical analysis to find a relationship between two different data sets of different granularities: atmospheric particles (and their elemental constituents) and elemental carbon (EC). Specifically, we wish to determine which elements in the atmosphere cause elemental carbon which is common in industrial zones and large cities and can normally be found in exhaust fumes and areas where there is visible carbon. In order to do this, we used machine learning regression algorithms including SVM regression and Lasso regression as well as regular linear regression.

BACKGROUND

Datasets:

ATOFMS (Atmospheric Time-Of-Flight Mass Spectrometer) Data

- Measurements of the amounts of different elements that make up an aerosol particle
- Data is visualized in a spectrum, where quantities of elements are represented as peaks



EC (Elemental Carbon) Data

- The amount of EC present in aerosol particles over a given period of time
- Generated by a machine that uses a filter substrate to collect aerosol particles, and then measures total EC using thermal methods within all particles captured

Goal:

- To discover a relation between aggregated ATOFMS data and EC.
- An equation that represents EC in terms of the most important peaks in ATOFMS output.

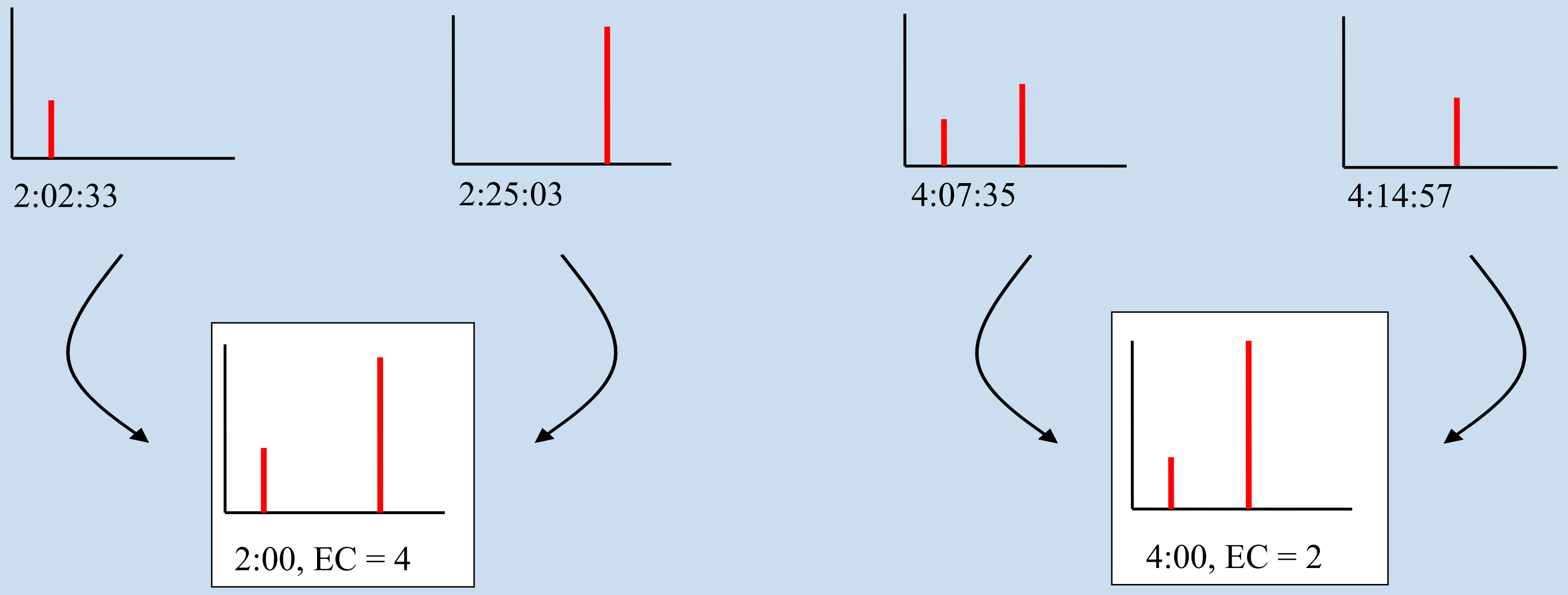
AGGREGATING THE DATA

Datasets are different granularities

- ATOFMS data collected as many as several times per second
- EC data collected on an hourly basis

Aggregation

- Round the time stamps of the ATOFMS particle data to the nearest hour
- Adjust the peak scale for the size of the particle
- Add together the peaks for each element of all the particles in each time bin
- Match the new aggregated ATOFMS data for each hour with the EC value for each hour



WEKA (Waikato Environment for Knowledge Analysis)

- Free package of machine-learning algorithms
- Variety of algorithms used for creating prediction models
- Popular among members of the data mining community

WEKA ALGORITHMS

SMOreg (Sequential Minimal Optimization)

- Used to train a support vector regression model
- Support vector machines are fast (much faster than a standard linear regression) and generally provide little error in regression models.

SMOreg with RBF (Radial Basis Function)

- Similar to SMOreg, except it is non-linear and uses an RBF kernel instead of a polynomial one
- Tends to be more accurate than regular SMO regression.

PROBLEMS WITH WEKA

Most important peaks

- Higher coefficients do not necessarily identify the most important peaks
- SVM algorithms are susceptible to "splitting" the importance of a particular element between multiple elements that represent roughly the same information.

Linear regression with multiples of 12

- Linear regression using elements whose atomic masses are multiples of 12 (C, C2, C3, etc…) performs just as well as SMO regression with all elements available for use

Too many variables

- A model with only a few elements is preferable to a model with many elements because it is more likely to be applicable to many situations

R

- Statistical analysis language
- Provides a different set of algorithms than WEKA

R ALGORITHMS

Lasso Regression

- Reduces as many of the coefficients as possible to zero
- A parameter that we pick, called the "bound," drives how many features are left
- By reducing the amount of features that are used in the model, lasso isolates the more important elements

Least Squares Regression

- Ran least squares regression on only the elements chosen by Lasso regression to get a model
- Leave-one-out cross validation used to determine prediction accuracy.

WOODSMOKE

- Particles containing woodsmoke obscure the general model for EC
- Solution is to make two models-one for low woodsmoke and one for high woodsmoke
- Separate particles into low and high woodsmoke based on the "angstrom coefficent"
- The data was difficult to cleanly separate into two pieces.

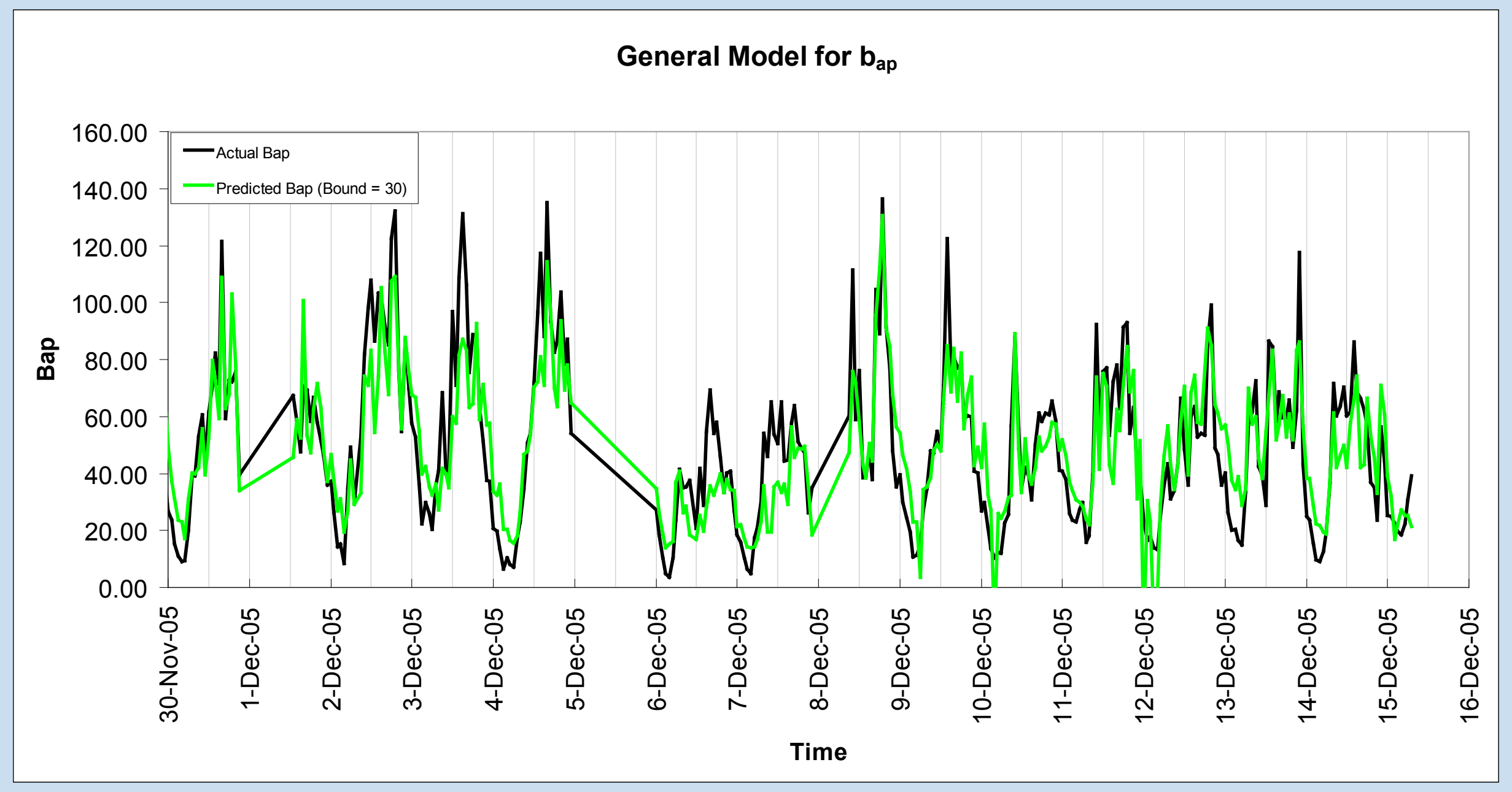
ABSORPTION COEFFICIENT

- The absorption coefficient (b_{ap}) is a value obtained through measuring the light absorbed by particles.
- EC values can be obtained from this quantity through the use of a conversion factor.
- In general, this conversion factor is not consistent.
- We provide an example model of b_{ap} prediction here.

MODELS

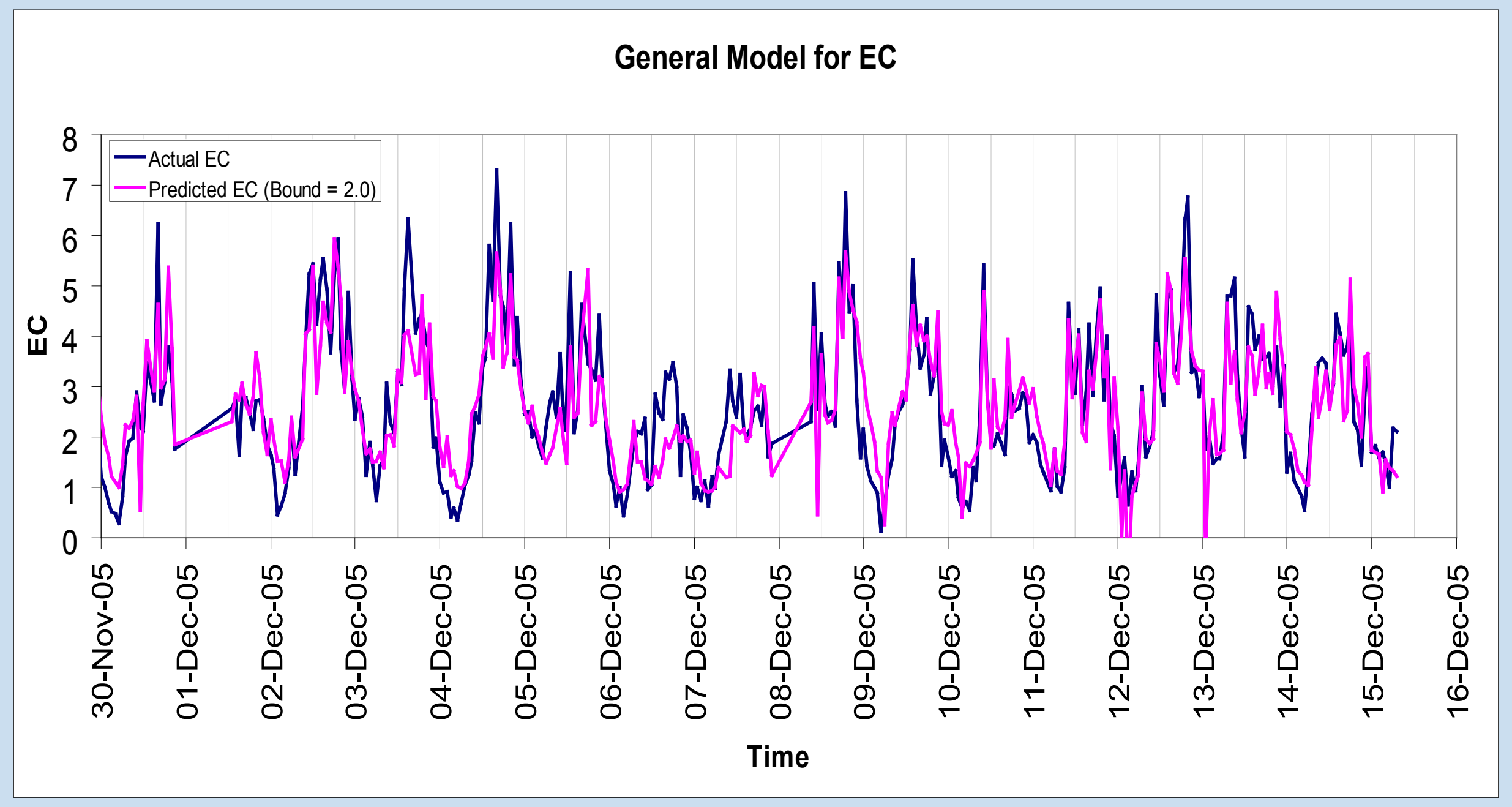
Correlation Coefficient: 0.8456
Mean Absolute Error: 11.7485

Attributes	Coefficients
(Intercept)	11.6125
mz-48	0.0144
mz-97	0.0055
mz84	0.0564
mz40	0.0008
mz-210	-0.1460
mz112	-0.1215
mz-131	-0.0448
mz286	-0.2252
mz215	0.0598
mz-93	0.0082
mz213	-0.0095
mz-153	0.0128
mz36	0.0007
mz-225	-0.0247
mz48	-0.0016



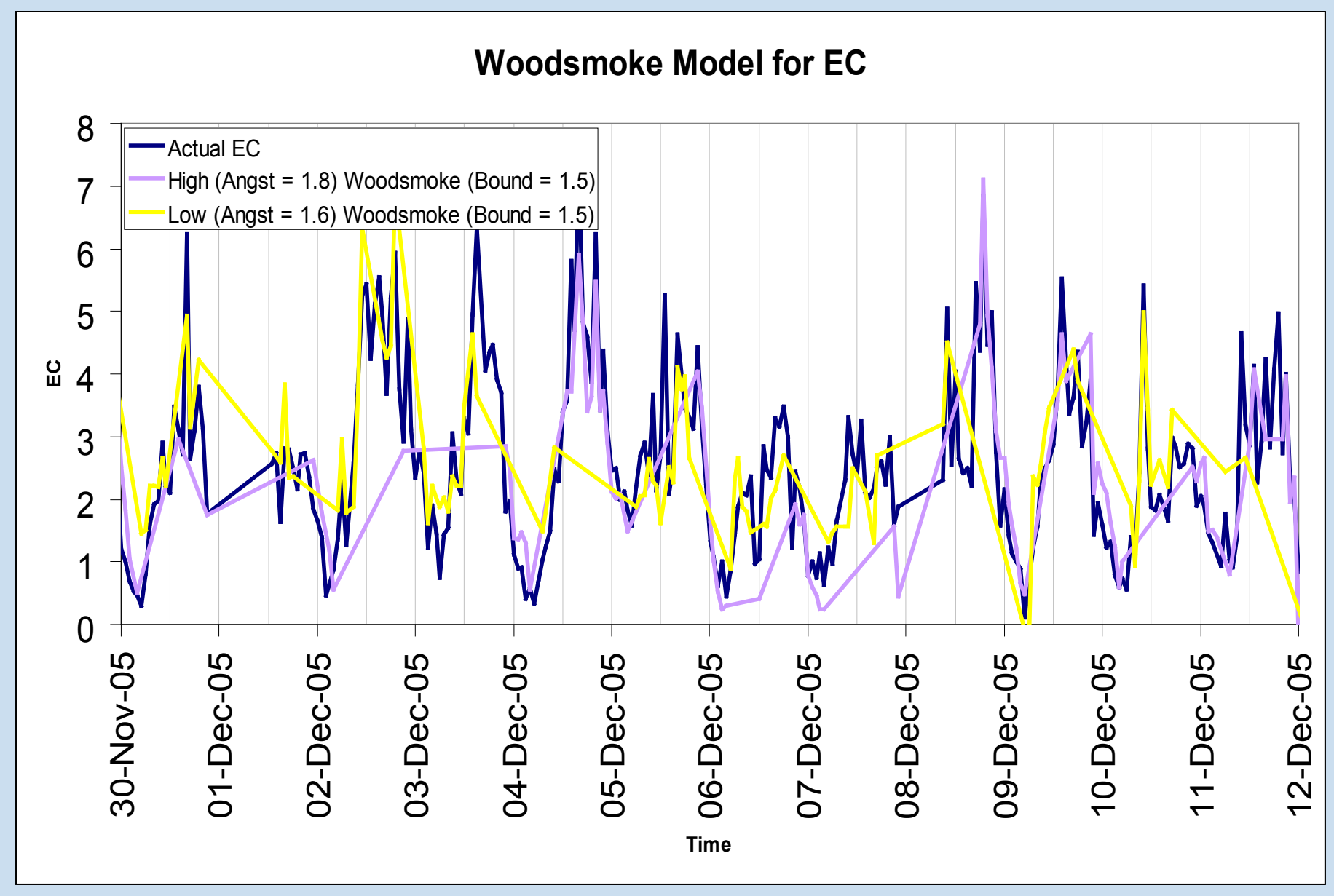
Correlation Coefficient: 0.8213
Mean Absolute Error: .6082

Attributes	Coefficients
(Intercept)	0.8205
mz-48	0.0006
mz127	0.0069
mz221	-0.0156
mz230	0.0089
mz-52	-0.0072
mz-61	-0.0006
mz-50	0.0010
mz112	-0.0061
mz-93	0.0004
mz55	0.0000
mz-104	-0.0025
mz12	0.0003
mz-66	0.0015
mz20	0.0069
mz-254	-0.0120



Correlation Coefficient: .8046
Mean Absolute Error: 0.5697

Attributes	Coefficients
(Intercept)	1.1983
mz-299	0.1205
mz-2	0.0033
mz-253	-0.0297
mz154	-0.0022
mz207	-0.0008
mz105	-0.0075
mz-48	0.0002
mz84	0.0034
mz-109	0.0015
mz-102	-0.0060
mz-275	0.0831
mz157	-0.0029
mz74	-0.0013
mz87	0.0044
mz40	0.0000
mz76	0.0083



Correlation Coefficient: 0.9042
Mean Absolute Error: 0.5751

Attributes	Coefficients
(Intercept)	0.1311
mz-48	0.0014
mz128	0.0024
mz-153	0.0042
mz-15	-0.0044
mz-184	-0.0072
mz37	0.0003
mz213	-0.0004
mz-190	-0.0032
mz-11	-0.0291
mz50	-0.0001

CONCLUSIONS AND ONGOING WORK

We came up with several models that correlate quantities of different types of particles to EC. The correlation coefficients that we were able to obtain are looking very promising. Though the data is still being investigated, this suggests that we can in fact use the m/z values provided by an ATOFMS to better understand quantitative EC data, to the point where we can make accurate predictions about it. Carbon-4 (m/z -48) apppears to be an important contributer to EC as it is consistently one of the most important attributes (often the most important one) in the various models that we came up with. Additionally, we were able to find that we can achieve significantly better results when we allow the model to take any of the m/z values into account, as opposed to simply restricting it to m/z values that are multiples of 12 (Carbon). Finally, with regards to the woodsmoke data, although it is difficult to accurately determine a cut-off point between high and low woodsmoke, we did manage to find some high correlation coefficients when we generated models for certain cutoff points. Since the models for each side of these cutoff points were very different, this suggests that woodsmoke has an effect on what types of particles are used to predict quantities of EC in the air.

ACKNOWLEDGEMENTS

Our work is funded by NSF ITR grant IIS-0326328. The atmospheric data sets were provided by Deborah Gross of the Carleton College Chemistry Department, and Jamie Schauer and Dave Snyder from the University of Wisconsin - Madison. The original version of the aggregation algorithm that we used (in many different ways, modifying it to fit our purposes) was written by Leah Steinberg.