Using New Statistical Approaches to Update Daily Ozone Concentration Forecasting Tools

Marcus Hylton, Nathan Pavlovic, Patrick Zahn
Sonoma Technology, Inc.
Petaluma, CA

for

National Air Quality Conference
Austin, Texas

January 25, 2018

Sonoma Technology, Inc.

Background – Dayton

- STI has provided tools to forecast daily ozone concentrations for the Regional Air Pollution Control Agency in Dayton, Ohio, since 2008
- Tools are typically updated with new air quality and meteorological data every 1 to 2 years, and evaluated for accuracy
- Goal is to predict when ozone levels will be Unhealthy for Sensitive Groups (USG) or higher (≥ 0.071 ppm for the daily maximum 8-hour ozone average)

RAPCA
Regional Air Pollution Control Agency

Background – AQCast

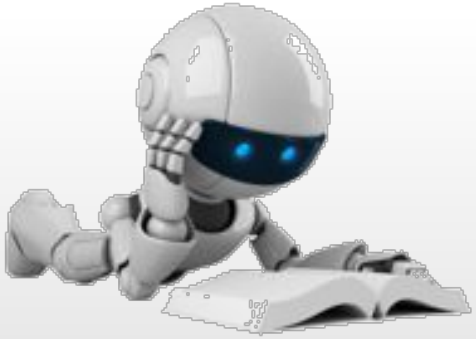
- Decision tool developed using observed pollutant concentrations and meteorological variables
- Automatically runs ozone and PM regression equations and Classification and Regression Tree (CART) daily
- Archives all forecasts and model data

AQCast > Query Results — AQCast Key

Region	City	Pollutant	Model	Method	8/30	8/31	9/1	9/2	9/3
Demo	Demo	8hrO3 (ppb)	0 GFS	Regression (Log) - GFS	59.1	86.1	114.4	116.3	108.5
Demo	Demo	8hrO3 (ppb)	0 NAM	Regression (Log) - NAM	60.1	83.3	88.7		
Demo	Demo	8hrO3 (ppb)	12 NAM	Regression (Log) - NAM	61.7	80.6	87.4	95.1	

Background – What is Machine Learning?

- "Machine learning allows software applications to become more accurate in predicting outcomes without being explicitly programmed."
- There are many different machine-learning algorithms.

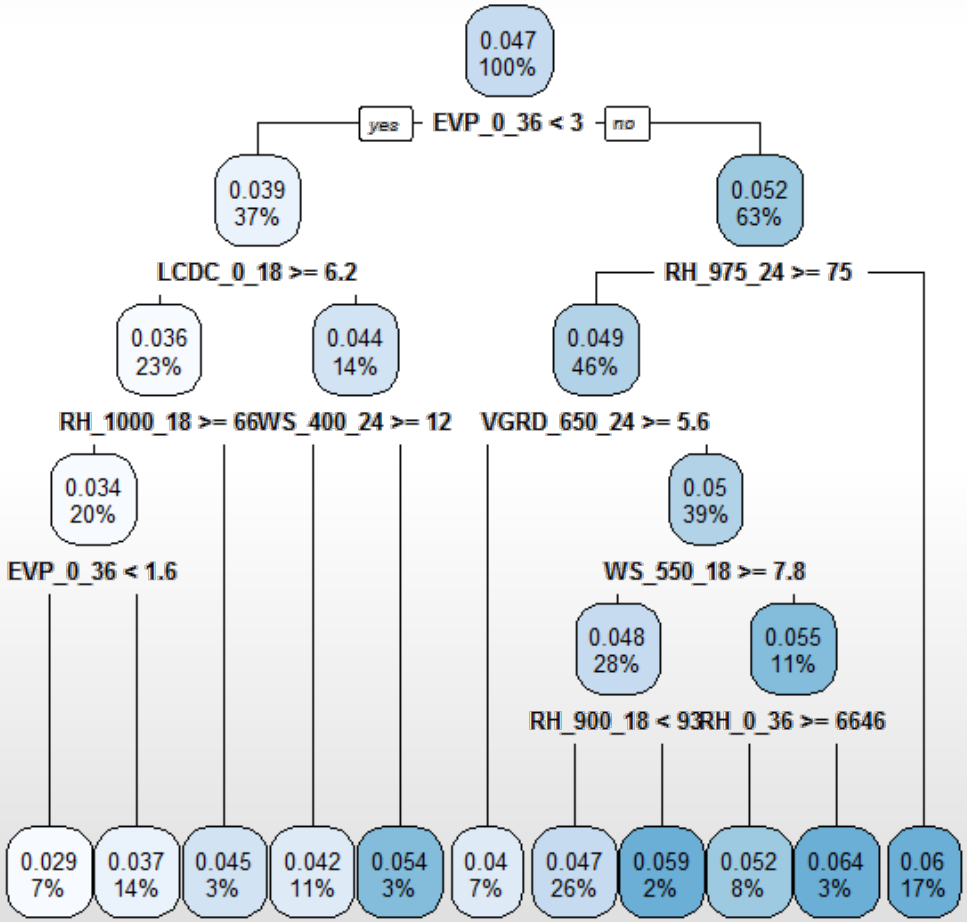


Classification & Regression Trees (CART)

- Predictions are split based on input variables
  - The number of splits and stopping rules are based on model input
- Produces end bins/nodes with a mean predicted value

Classification & Regression Trees (CART)

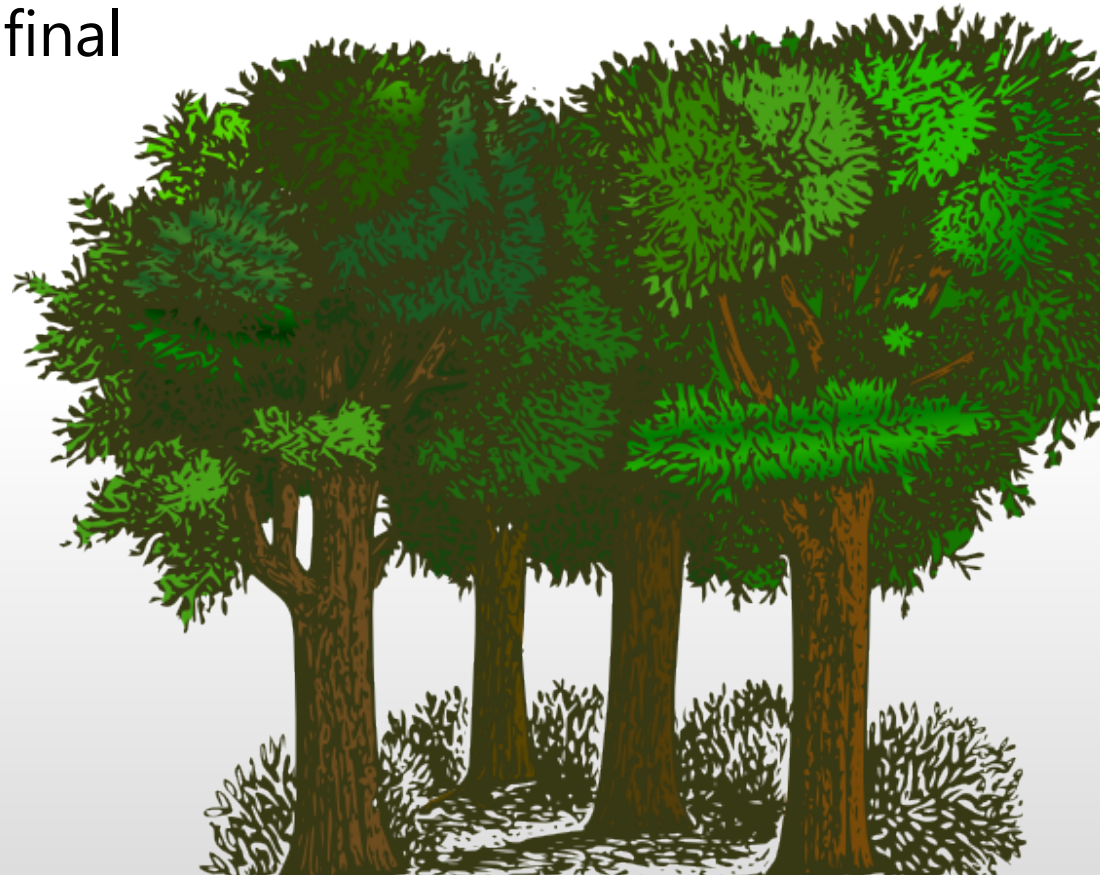
Example of CART decision tree



Random Forest

Ensemble of decision trees

- Results from all trees are combined to compute a final average prediction
- Known to be a fairly accurate predictive algorithm and is widely used



Extreme Gradient Boosting (XGBoost)

- Relatively shallow decision trees (few splits) are built iteratively
- Algorithm has been used to win a variety of Machine Learning competitions
- Disadvantage: higher effort and computational costs compared to some other models

dmlc
XGBoost

Statistical Measures

- R-Squared
  - Statistical measure of how close data are to a fitted regression line
  - Higher percentages are best
- Probability of Detection (POD)
  - Of all observed days above a threshold, POD shows the percent of days that the model's predicted conditions exceeded the threshold
  - Higher percentages are best
- False Alarm Rate (FAR)
  - Of all predicted days above a threshold, FAR shows the percent of days that the observed conditions did not exceed the threshold
  - Lower percentages are best

Previous Method of Development

- Developed the equations using observed meteorological parameters
  - Hourly data from surface stations
  - Soundings (limited to twice a day on most days)
- Data were compiled into a Microsoft Access database
  - We aggregated and calculated data to get it into a suitable format for comparison to model output data
- CART and Regression equations were developed in Systat 13
  - Required significant analyst input and trial/error
- **High POD (~75%) and reasonable FAR (~50%) on the USG threshold in training/testing, but when applied to the weather model data, our CART model performed poorly**

New Method of Development

- Develop the tools using *modeled* meteorological parameters rather than observed parameters
  - Use model GRIB files
  - Derive parameters from model data (e.g., temperature difference, recirculation)
- Use R to gather input data and train and test the model
  - CART (rpart & party packages)
  - Random Forest (randomForest package)
  - XGBoost (xgboost package)

Data Sources

- Global Forecast System (GFS) and North American Mesoscale Model (NAM) weather data
  - December 2013–June 2017
- AQS daily maximum 8-hour ozone concentrations for Dayton, OH
  - 4 monitoring sites
  - April 2014–June 2017

Year	Number of Exceedances
2014	3
2015	7
2016	9
Jan.–June 2017	5

Data Sets

- Training
  - 75% of days during the ozone season in 2014, 2015, and 2016
- Testing
  - 25% of days during the ozone season in 2014, 2015, and 2016
- Validation
  - April through June 2017

Methods – NAM Model Predictors

57 NAM parameters + derived parameters + yesterday's ozone observations

20 levels (for certain parameters)

4 forecasted values throughout the day

=

756 predictor variables for a given day

NAM CART – Initial Runs

- 35 models
  - No bins predicted a value above 0.07 ppm
  - Highest R-squared value was 0.45
- Possible ways to improve the model
  - Up-sampling: Puts more weight on high impact/USG days
  - Do not include days under a certain ozone concentration threshold

Model Tweaking/Adjusting

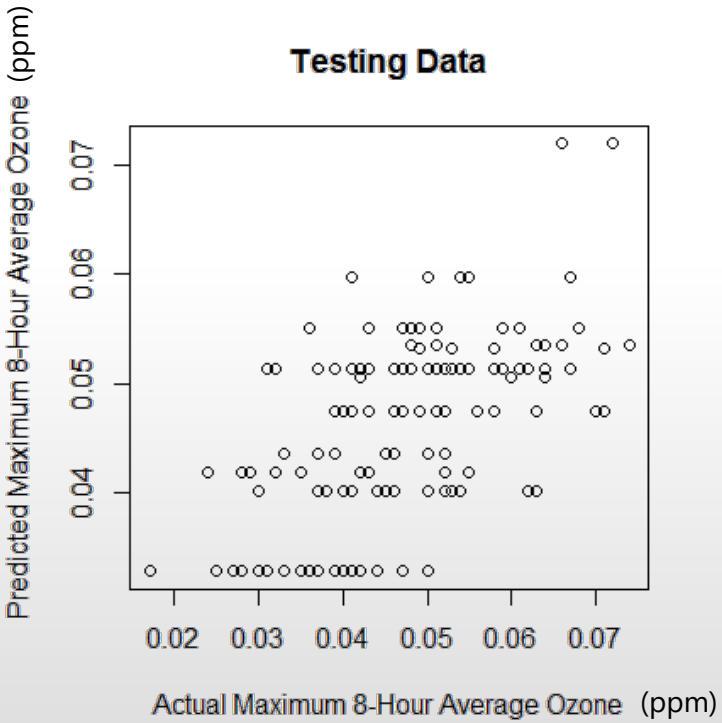
Up-sampling only

- Sweet spot around an up-sample rate is between 10 and 16 times

o3t	upr	pod	far	rsq	pod_train	far_train	rsq_train
0	1	0		0.358493	0		0.470039
0	2	0	1	0.37165	0.1	0.285714	0.505434
0	4	0		0.372606	0		0.596345
0	6	0	1	0.349648	0.32	0.428571	0.569041
0	8	0		0.309271	0		0.560456
0	10	0.6	0.625	0.352986	0.72	0.357143	0.568711
0	12	0.75	0.625	0.364163	0.9	0.357143	0.554942
0	14	0.6	0.65625	0.387828	0.72	0.392857	0.523019
0	16	0.75	0.65	0.397134	0.9	0.385714	0.534353
0	18	0.45	0.75	0.349091	0.78	0.561265	0.487711
0	20	0.5	0.75	0.35611	1	0.545455	0.498781
0	22	0.5	0.75	0.348254	1	0.545455	0.462052
0	24	0.4	0.8	0.354117	0.92	0.544056	0.477225
0	26	0.5	0.75	0.35903	1	0.545455	0.465461
0	28	0.5	0.75	0.358998	1	0.545455	0.465267
0	30	0.5	0.75	0.358972	1	0.545455	0.465115

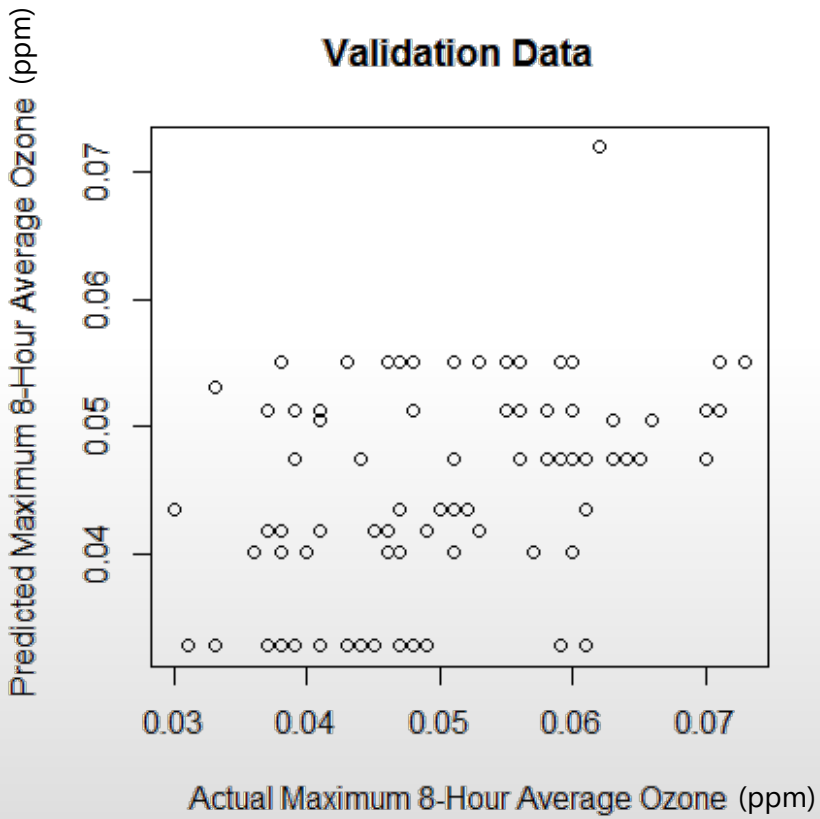
Final NAM Model – Testing/Training

Type	O_3 Threshold	Up-Sample Rate	Probability of Detection	False Alarm Rate	R-squared
Testing	0	14	0.25	0.5	0.308
Training	0	14	1	0.23	0.565



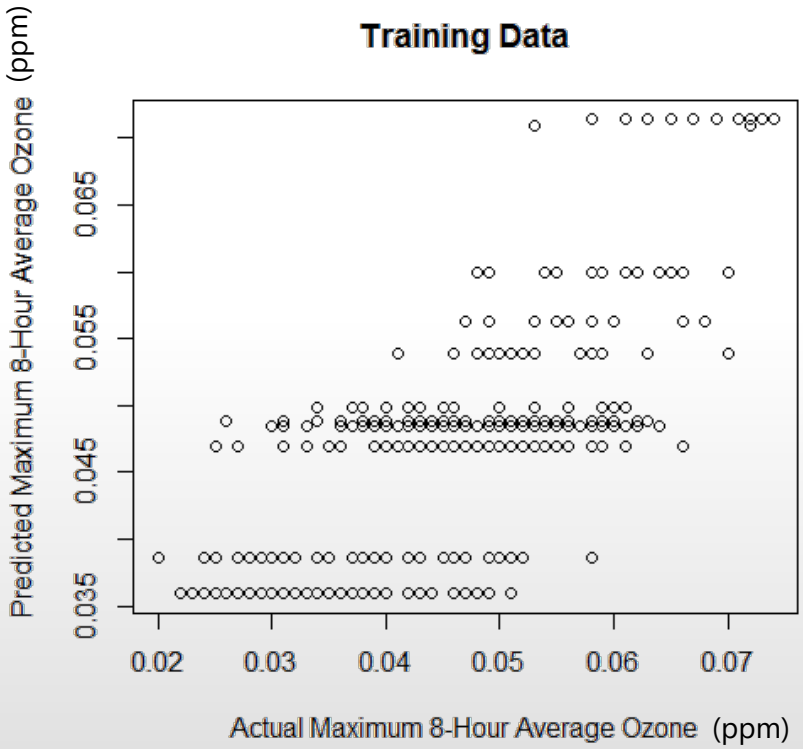
Final NAM Model Validation Results – 2017

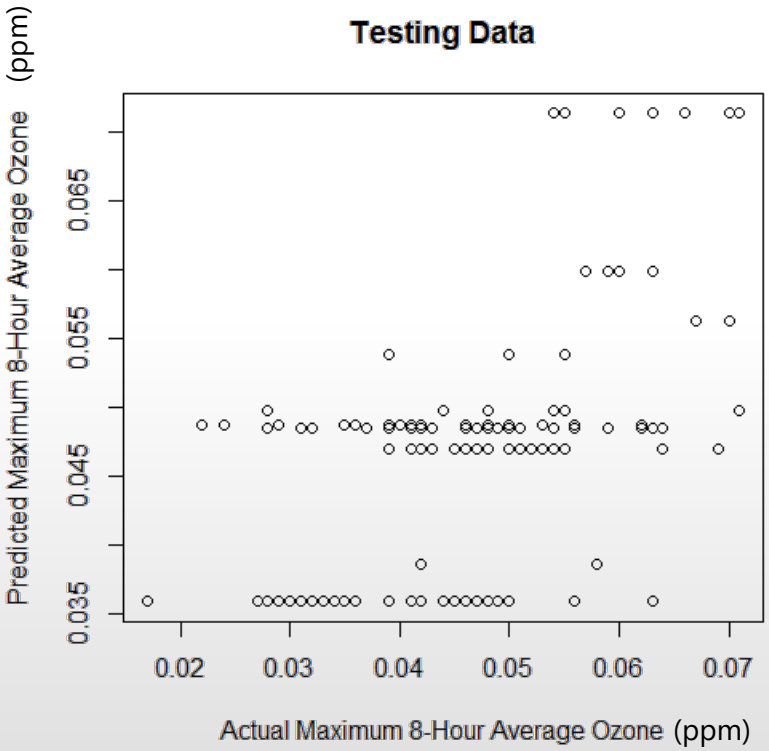
Type	O_3 Threshold	Up-Sample Rate	Probability of Detection	False Alarm Rate	R-squared
Validation	0	14	0 (0 of 3)	1 (1 of 1)	0.175



Final GFS Model – Testing/Training

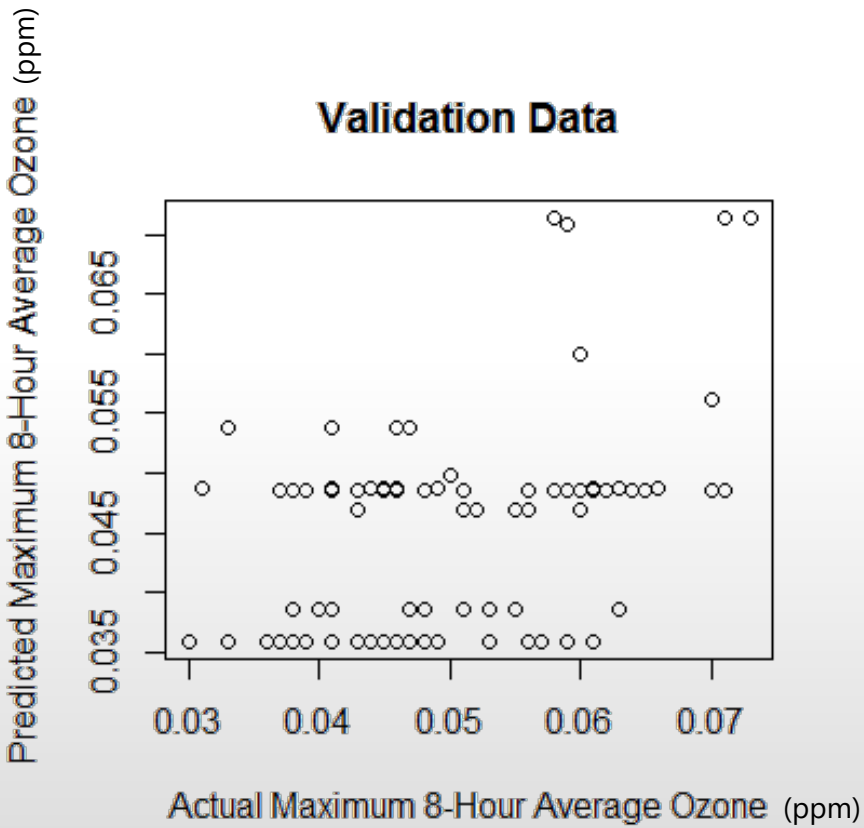
Type	O_3 Threshold	Up-Sample Rate	Probability of Detection	False Alarm Rate	R-squared
Testing	0	14	0.5	0.889	0.301
Training	0	14	0.9	0.4375	0.473



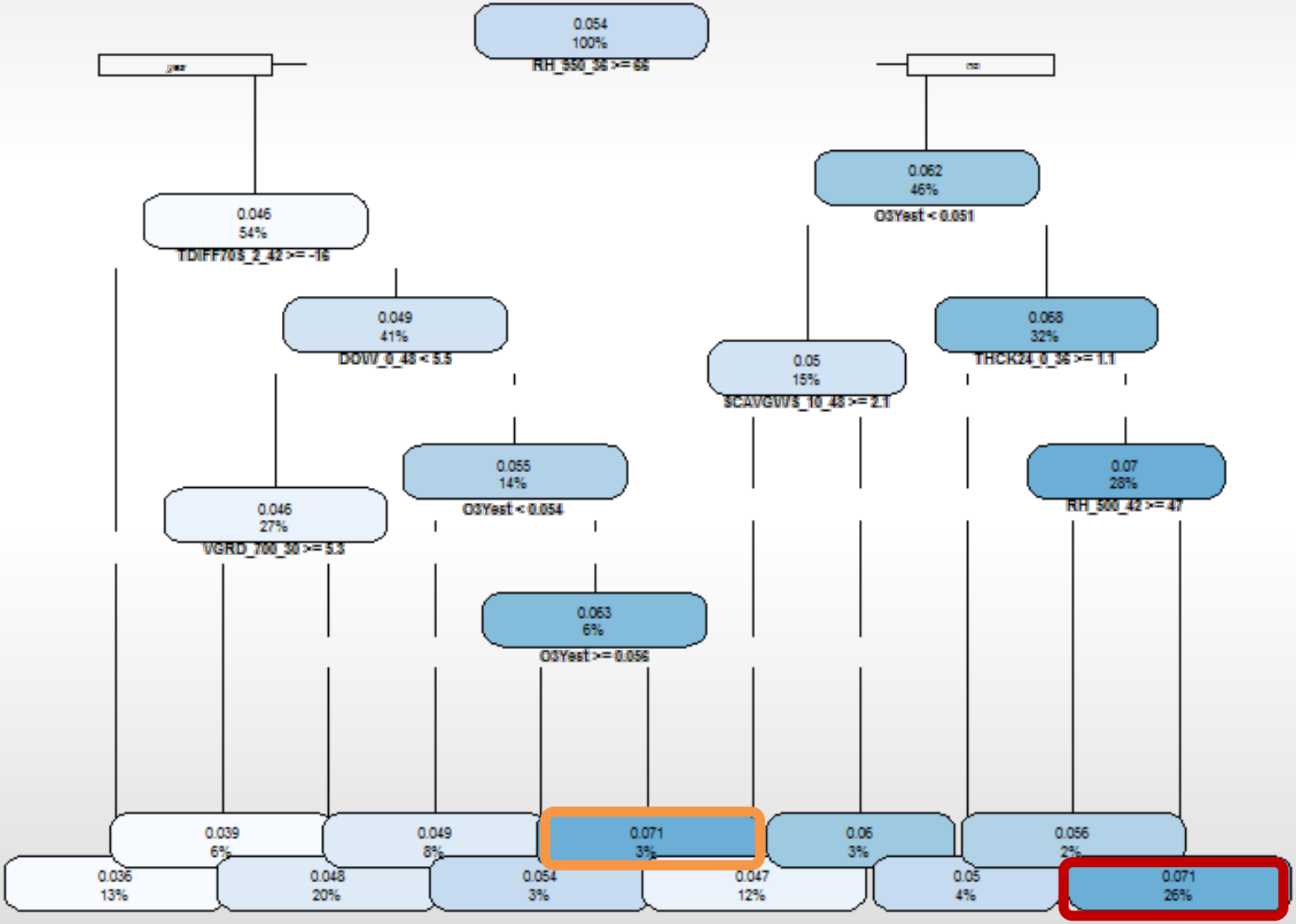


Final GFS Model – 2017 Validation

Type	O_3 Threshold	Up-Sample Rate	Probability of Detection	False Alarm Rate	R-squared
Validation	0	14	0.6 (2 of 3)	0.5 (2 of 4)	0.187



Final GFS Model



Final GFS Model

USG Bin 1*

Predicted value: 71 ppb
3% of days in the training set
fell in this category

Variables:

- Relative Humidity at 950 mb at hour 36 is ≥ 66%
- Temperature Difference 700 mb to Surface is < -16°C
- Day of the week is > 5.5 (Saturday or Sunday)
- Yesterday's Ozone is ≥ 54 ppb and < 56 ppb

USG Bin 2

Predicted value: 71 ppb
26% of days in the training set
fell in this category

Variables:

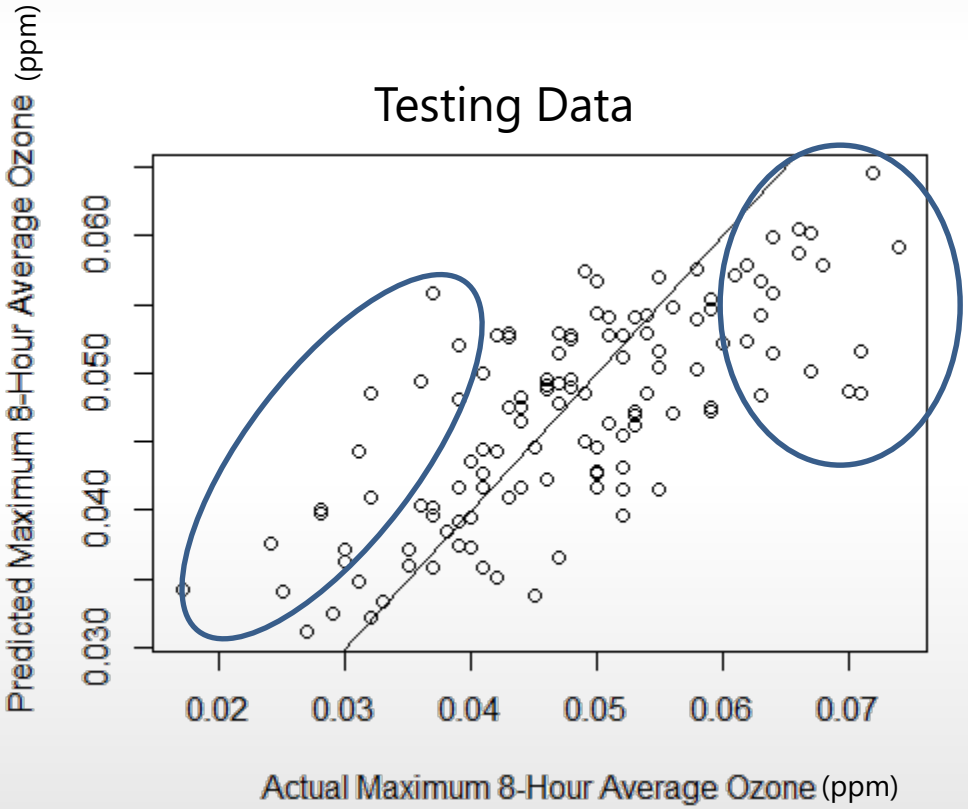
- Relative Humidity at 950 mb at hour 42 is < 66%
- Yesterday's Ozone is ≥ 51 ppb
- 24-hour thickness difference between 1000-500 mb at hour 36 is < 1.1 m
- Relative Humidity at 500 mb at hour 42 is < 47%

*Outlier USG day

Dayton – Random Forest Regression NAM Results

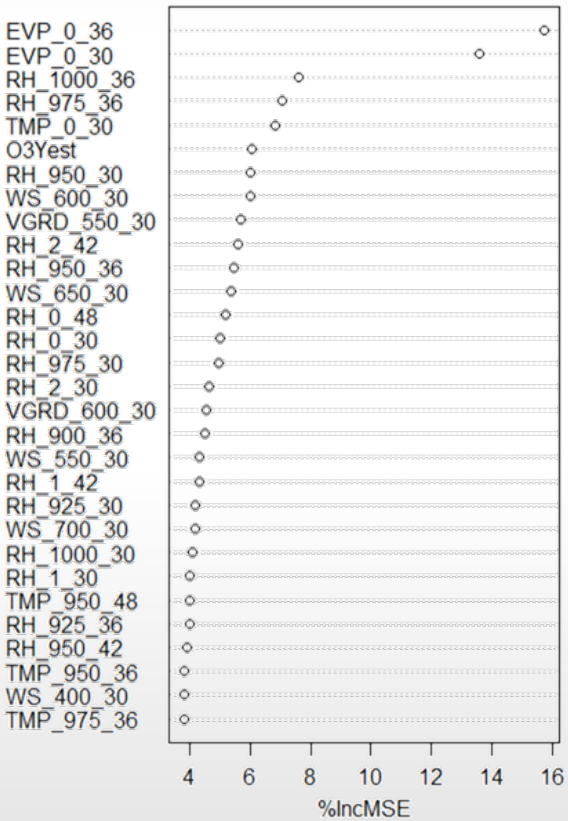
High values were underestimated and low values were overestimated

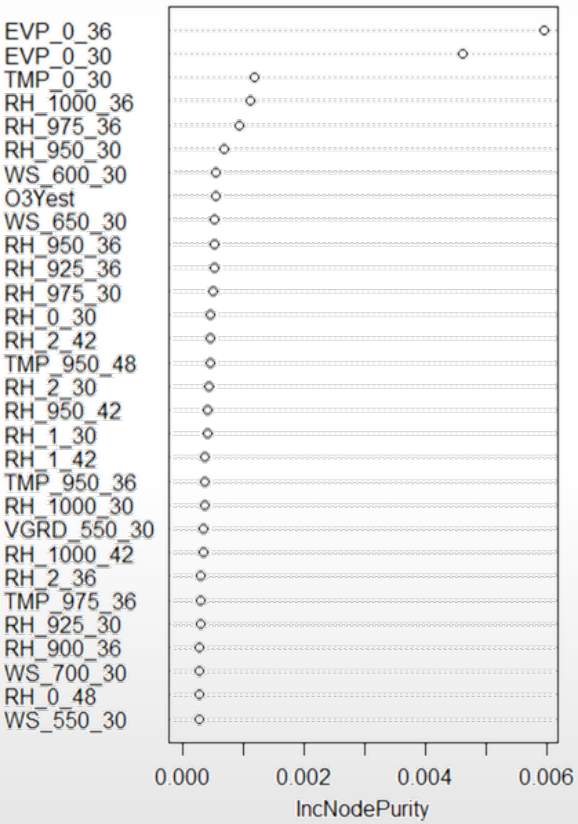
- Typical for Random Forest
- For our purposes (predicting the few high ozone days), this model did not perform well



Most Important Variables – NAM Regression Random Forest

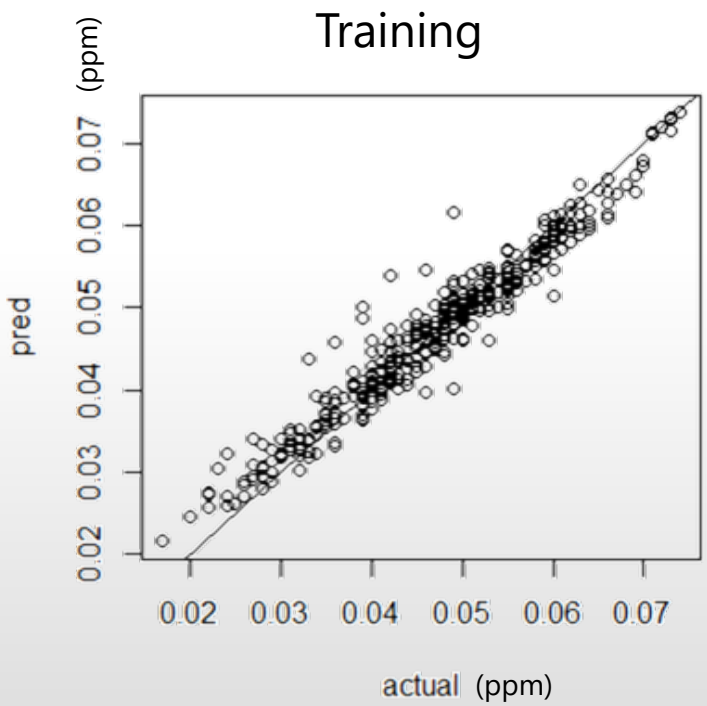
- Surface evaporation (+)
- Low-level relative humidity (-)
- Surface Temperature (+)

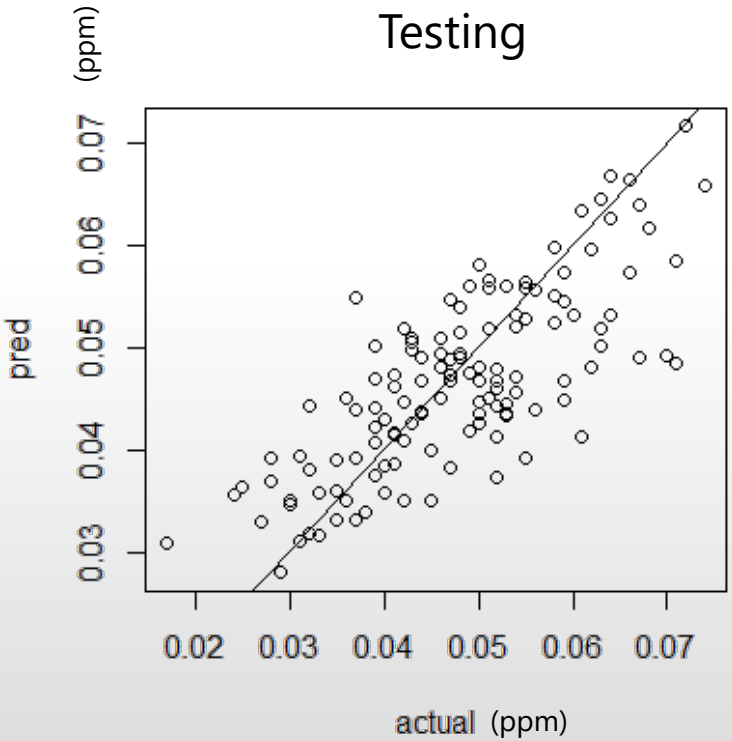




Dayton – XGBoost Regression NAM Results

Type	Probability of Detection	False Alarm Rate	R-squared
Testing	0.25	0	0.578
Training	0.928	0	0.972
Validation	0	N/A	0.49





Assessing the Viability of the Models

- To build a reasonable tool, a certain percent of days should be above the chosen threshold
  - Lower the ozone threshold for prediction (e.g., 65 ppb instead of 70 ppb) for Dayton

  Or
  - Develop and test new models on a city that has a larger percentage of USG days and see how it performs
- We chose Sacramento, CA, to see how the models performed
  - 14x up-sample rate retained

Year	Number of Exceedances
2014	38
2015	20
2016	33
2017	12

Sacramento Results

Validation Data Sets (2017)

Type	Model	Probability of Detection	False Alarm Rate	R-squared
CART	GFS	0.89	0.69	0.49
CART	NAM	0.83	0.81	0.40
Random Forest Regression	NAM	0.33	0.45	0.63
Random Forest Regression	GFS	0.44	0.38	0.60
XGBoost Regression	NAM	0.56	0.50	0.59
XGBoost Regression	GFS	0.61	0.52	0.64

Pros of New Method of Development

- Apples to apples (model vs. model accounts for model bias)
- More variables to train the model on
  - 700+ variables per model
- Faster development and more fine tuning
- Runs through many more iterations than previous tool developments
- Options on what machine learning method to use

Cons of New Model

- Increased computational requirement for training the models
- Learning curve of which parameters to adjust in the model
- Equations/CARTs are unique by model type and model run
  - Unlike using observed conditions, a single point in time can have multiple predicted values (one for each model initialization)
  - Weather models each have their own quirks and biases, and applying one weather model's developed equation to another weather model (e.g., NAM Cart to GFS) would reduce accuracy

Future Ideas for Improvement

- Modify machine learning parameters
- Consider impacts outside of modelled parameters
  - Removing smoke days would remove several of the USG days from model consideration, but may improve performance
  - Holidays or event days (parades/concerts/fireworks, etc.)
- More years of weather and ozone data will improve the model
  - This will happen over time as we continue to add model data into our database

Contacts

Marcus Hylton
Meteorologist
mhylton@sonomatech.com
707.665.9900

Nathan Pavlovic
Air Quality Scientist
npavlovic@sonomatech.com
707.665.9900

Patrick Zahn
Meteorologist / Lead Forecaster
pzahn@sonomatech.com
707.665.9900

Sonoma Technology, Inc.
Environmental Science and Innovative Solutions

sonomatech.com @sonoma_tech