Data Driven Ozone Forecasting in Texas Urban Areas

Atmospheric and Environmental Research

Richard Pernak, Matthew Alvarado (malvarad@aer.com), Jennifer Hegarty, and Thomas Nehrkorn

Introduction

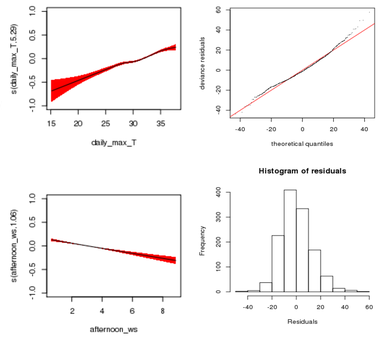
- We developed three statistical models that can forecast tomorrow's MDA8 O_3 in six urban areas in Texas.
  - Quantitative forecasts using generalized additive models (GAMs)
  - Probabilistic forecasts using GAMs and logistic regression
  - Classification forecasts using the Random Forest method
- Here we show how these forecast models performed during the 2016 O_3 season and the improvements made to each model based on this evaluation.

Generalized Additive Models (GAMs)

- GAMs are extensions of multiple linear regression models that fit unknown non-linear functions of predictors

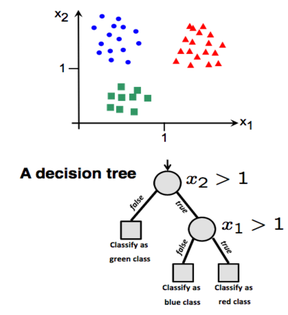
$$g(\mu_i) = \beta_o + f_1(x_{i,1}) + f_2(x_{i,2}) + \cdots f_n(x_{i,n}) + f_p(D_i) + W_d$$

- Allows quantitative predictions of tomorrow's MDA8 O_3.
- Can also predict probability of tomorrow's MDA8 O_3 exceeding a given threshold using logistic regression.



Random Forest

- Decision tree models can be constructed to predict tomorrow's ozone air quality class in terms of meteorological criteria.
- However, they tend to over-fit the data.
- Random forests average different decision trees together to reduce this over-fitting.



Near-Real-Time Data Sources

- Meteorological Forecast Data
  - Model Output Statistics (MOS) forecasts based on the numerical North American Mesoscale 12-km (NAM-12) forecasts from NCEP.
  - NAM-12 upper air temperatures and geopotential heights.
- TCEQ Monitor Ozone Data

Table 1. TCEQ and MOS sites corresponding to the six urban sites investigated in this report.

TCEQ Site Abbreviation	MOS Site Abbreviation	Associated City
ARR	KAUS	Austin/Round Rock
BPA	KBPT	Beaumont/Port Arthur
DFW	KDFW	Dallas/Fort Worth
HGB	KIAH	Houston/Galveston/Brazoria
SA	KSAT	San Antonio
TLM	KGGG	Tyler/Longview/Marshall

Predictor Variables Used in the Forecast Models

Significance codes are: '***', significant at the α = 0.001 level; '**', significant at the α = 0.01 level; '*', significant at the α = 0.05 level; '.', significant at the α = 0.1 level.

Version 1.3

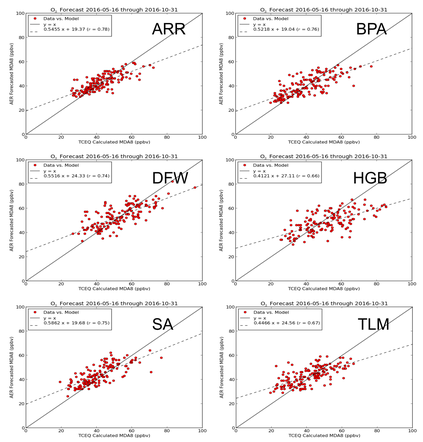
	ARR	BPA	DFW	HGB	SA	TLM
Day of Week	***	***	***	***	***	***
MOS.Afternoon.T	***	***	**	***	**	*
MOS.Diurnal.T	***	**		***	*	*
MOS.Mean.RH			***		***	
MOS.Mean.Dew.Point	***	*	*	***	*	
MOS.Mean.Wind.Speed	***	***	***	***	***	***
MOS.Mean.Wind.Direction	***	***	*	***	**	***
NAM_T_Diff (925 mb)						
Day of Year	***	***	***	***	***	**
MDA8.O3.max	***	***	***	***	***	***

Version 2.0

Predictor	ARR	BPA	DFW	HGB	SA	TLM
Day of Week	***	***	***	***	***	***
Day of Year	***	***	***	***	***	***
Previous Day MDA8	***	***	***	***	***	***
MOS.Afternoon.T	***	***	***	***	*	***
MOS.Diurnal.T	.	.	***	*	***	*
H_2O in g m^{-3}	***	***	***	***	***	***
MOS.Mean.Wind.Speed	***	***	***	***	***	***
MOS.Mean.Wind.Direction	*	***	.	***	.	

Quantitative Forecasts

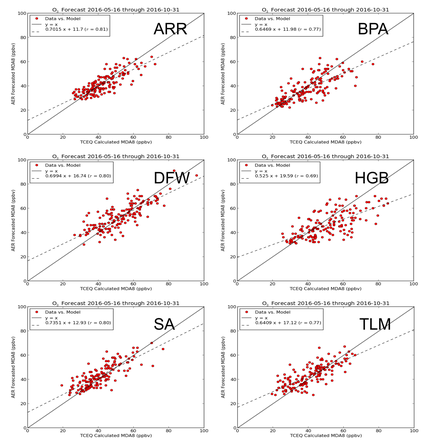
Version 1.3



Best-fit line slope (a) and intercept (b) and Pearson correlation coefficient (r). The mean (μ) and standard deviation (σ) of the model-data residuals is also provided.

Urban Area	*a*	*b*	*r*	*μ (ppbv)*	*σ (ppbv)*
ARR	0.55	19.37	0.78	-0.07	6.12
BPA	0.52	19.04	0.76	-0.09	7.91
DFW	0.55	24.33	0.74	0.63	8.28
HGB	0.41	27.11	0.66	-2.99	10.36
SA	0.59	19.68	0.75	2.65	6.98
TLM	0.45	24.56	0.67	1.61	8.19

Version 2.0

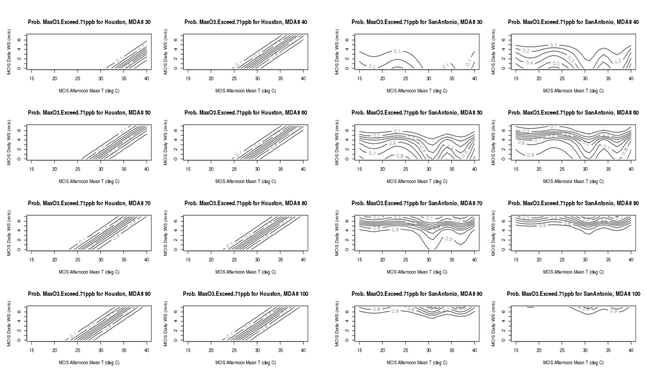


As on left but for the v2.0 models.

Urban Area	*a*	*b*	*r*	*μ (ppbv)*	*σ (ppbv)*
ARR	0.70	11.70	0.81	-1.04	5.74
BPA	0.65	11.98	0.77	-2.14	7.80
DFW	0.70	16.74	0.80	0.90	7.54
HGB	0.53	19.59	0.69	-4.71	9.98
SA	0.74	12.93	0.80	2.04	6.42
TLM	0.64	17.12	0.77	2.27	7.02

Probabilistic Forecasts

HGB, Prob > 71 ppbv **SA, Prob > 71 ppbv**



Reliability and ROC Area statistics for the v1.3 model of each urban area.

Urban Area	*Threshold (ppb)*	*Reliability*	*ROC Area*
ARR	55	0.00	0.90
BPA	55	0.00	0.79
DFW	55	0.04	0.88
DFW	71	0.01	0.91
HGB	55	0.10	0.51
HGB	71	0.00	0.80
SA	55	0.01	0.89
TLM	55	0.15	0.88
TLM	71	0.01	0.99

Reliability and ROC Area statistics for the v2.0 model of each urban area.

Urban Area	*Threshold (ppb)*	*Reliability*	*ROC Area*
ARR	55	0.01	0.89
BPA	55	0.01	0.78
DFW	55	0.00	0.90
DFW	71	0.01	0.92
HGB	55	0.02	0.82
HGB	71	0.00	0.78
SA	55	0.01	0.91
TLM	55	0.03	0.92
TLM	71	0.01	0.99

Classification Forecasts

MDA8 classes used in the Version 2.0 random forest model.

Color Class	O_3 MDA8 Range (ppb)
Green (Good)	MDA8 < 55
Yellow (Moderate)	55 ≤ MDA8 < 71
Orange (Unhealthy for Sensitive Groups)	71 ≤ MDA8 < 86
Red (Unhealthy)	86 ≤ MDA8 < 105
Purple (Very Unhealthy)	105 < MDA8

Classification forecast success, false alarm, and miss rates for all models.

Urban Area	*Success Rate*	*False Alarm Rate*	*Miss Rate*
Austin	87.01%	4.54%	8.44%
Beaumont	86.36%	3.25%	10.39%
Dallas	77.27%	7.14%	15.58%
Houston	66.88%	1.95%	31.17%
San Antonio	89.12%	2.72%	8.16%
Tyler	90.26%	2.60%	7.14%

Conclusions

- The v2.0 quantitative forecast models are a significant improvement over v1.3, with higher correlations, better slopes, and less variance.
- The v2.0 quantitative forecasts explain over 70% of the deviance (variance) in each urban area, a significant improvement upon the v1.3 models (56-62%).
- The probabilistic models show little predictive skill.
- The random forest classification models are able to correctly predict the AQI classification 67% of the time in HGB, 77% of the time in DFW, and over 86% of the time elsewhere.
- False negatives were more likely than false positives.
- We recommend that a combination of the quantitative forecasts and the classification forecasts be used.
- Future work will focus on improving the performance of the forecasts in the high tail of the observations to reduce the "miss rate" where the severity of a poor ozone event is under-predicted.

A PORTION OF THE DATA OR OTHER INFORMATION IN THIS DOCUMENT WAS PREPARED UNDER A CONTRACT FROM THE TEXAS COMMISSION ON ENVIRONMENTAL QUALITY (TCEQ)
The contents of this document do not necessarily reflect the views and policies of the TCEQ, nor does the TCEQ endorse any trade names or recommend the use of any commercial products mentioned in this document.