Transactions on Ecology and the Environment vol 29 © 1999 WIT Press, www.witpress.com, ISSN 1743-3541

A semi-empirical model relating meteorology and ozone in the San Francisco Bay Area

Lawrence C. Larsen
State of California Air Resources Board, Planning and Technical Support Division, 2020 L Street, Sacramento, California, USA
Email: llarsen@arb.ca.gov

Abstract

Recent ozone trends in the San Francisco Bay Area have been volatile, with a total of 13 days exceeding the national 1-hour ozone standard of 124 ppb in 1995, 12 in 1996, and 8 in 1998 but only 3days in 1993, 2 in 1994, and 0 in 1997. The specific causes of this volatility are elusive due to the complex chemical reactions and variable meteorology that govern tropospheric ozone; these factors also make it difficult to determine the effects of reducing the emissions of ozone precursors. To help understand ozone trends, this study develops a semi-empirical model relating daily ozone to daily meteorological conditions in the Bay Area. Meteorological parameters used in the model include pressure and temperature gradients, mixing height, temperature (surface and aloft), cloud cover, and wind speed. An indicator for weekend days is also included. The model is constructed in two stages. First, synoptic-scale pressure gradients and meso-scale temperature gradients, summarized by principal components, are used to cluster all days in the May – October ozone seasons for 1988 – 1996. Second, within-cluster additive models use other variables to quantify daily basin-wide maximum ozone concentrations. The within-cluster models are calibrated using daily ozone data from 1993 through 1995 only. The overall model's performance is very encouraging, achieving an R^2 of 83 percent for the 1993 – 1995 data.

The modeling approach presented in this study uses meteorological data that are available routinely in many regions, suggesting that the method may be an effective tool in other regions.

1 Introduction

Tropospheric ozone is the chief constituent of urban smog and control of this pollutant continues to be a significant challenge. Although the emissions of ozone precursors have been substantially reduced in many areas, variable meteorological conditions can mask the effects of the reductions. Twenty years ago, Davidson[1] and Zeldin and Meisel[2] were among the pioneers who sought effective methods of removing meteorological variability from analyses of ozone trends. More recently, alternative methods have been developed by Cox and Chu[3], Rao and Zurbenko[4], and Davis, *et al.*[5] The work described in this paper builds on Davis' method with modifications motivated by scientific principles of atmospheric processes and ozone formation.

Davis, *et al.*, developed a two-step approach to the problem of relating ozone concentrations to meteorological conditions and applied their method to data around Houston, TX. In the first step of their procedure, a cluster analysis identifies groups of days that have similar synoptic weather conditions. In the second step, generalized additive models relate ozone to specific meteorological variables within the clusters.

Though similar to the methods used by Davis, *et al.*, the methods used in this study are somewhat simpler and, perhaps, more familiar to a wider audience. The reader is encouraged to refer to the paper by Davis, *et al.* for a general discussion of the principles.

2 Data

Routinely available sources supplied ozone data and meteorological data because a secondary goal of this research is to support future development of an automated system for computing meteorologically adjusted air quality trends over the past two decades.

2.1 Ozone data

Ozone data were extracted from the ADAM[6] data system at the California Air Resources Board for 28 sites in the San Francisco Bay Area for the period 1988-1996. In most cases, these data are identical to the data contained in the U.S. Environmental Protection Agency's AIRS data system though occasional differences may arise for various reasons.

For each site, values were extracted for the daily maximum ozone concentration and the hour at which this maximum was first reached. Only days for which the hourly data are considered complete were selected. Data were limited to an ozone season defined as May through October. Nine sites with less than 160 of 184 possible days in one or more years were discarded, leaving 19 sites that qualified for this study. Almost all of the discarded sites were highly incomplete. Two sites were retained even though their data were incomplete for a single year: Oakland-Alice St. (146 days in 1989) and Hayward-La Mesa (78 days in 1996). These two sites seldom record ozone concentrations that exceed the state ambient air quality standard for ozone of 90 ppb. Table 1 identifies the sites that provided ozone data and meteorology data.

Table 1. Stations contributing data for analyses

Type of data	Site ID Numbers				
Ozone data (AIRS #)	060010003	060010005	060011001	060012001	060130002
	060130003	060131002	060133001	060410001	060550003
	060750005	060811001	060850002	060850004	060851001
	060851002	060950002	060950004	060970003	
Meteorology (WMO #)	745060	745160	723890	724839	724940
	724945	724920	724880	725970	724930
Only surface temperature	Lawrence Livermore Labs				
	Fort Funston				

2.2 Meteorology data

Table 1 identifies the sites contributing meteorology data. The meteorology data were taken primarily from CD-ROM sources available from the National Climatic Data Center (NCDC). Upper-air data for Oakland in the center of the S.F. Bay Area came from two CDs of rawinsonde data covering the years 1988-1996. Most of the surface data were taken from a recent set of CDs from NCDC entitled *International Surface Weather Observations, 1982 – 1997*. Surface temperature data from two sites – Lawrence Livermore Labs and Fort Funston – were supplied by the staff of the Bay Area Air Quality Management District.

Table 2 identifies the variables used in the analyses. Synoptic-scale pressure gradients were based on sea-level pressures reported by NCDC. Three pressure gradients were defined as follows: Medford – Oakland (north); Reno – Oakland (east); and Fresno – Oakland (south). The pressure differences were determined circa 16:00 PST to coincide with the afternoon rawinsonde at Oakland.

Meso-scale temperature gradients reflect mean 10:00 – 16:00 PST surface temperatures at five sites; the data for two of these sites – San Francisco and Alameda – were averaged to represent the central Bay Area. Three temperature gradients were then defined as follows: Sacramento – Central Bay (northeast); Stockton – Central Bay (east); and San Jose – Central Bay (southeast).

Because non-straight-line relationships frequently prevail between ozone and surface temperature, a squared temperature variable was included. Similarly, an inverse wind-speed variable was included.

The wind speed vector was separated into a northern component vector (u) and an eastern component vector (v). The response of these components to the pressure and temperature gradients was an important cross-check for effectiveness of the clustering step in the model development.

The mixing heights at San Jose in the southern subregion of the Bay Area were based on the vertical temperature profile from the 4 a.m. PST rawinsonde at Oakland and the hourly surface temperature at San Jose at the first hour of the daily ozone maximum. All temperatures were converted to sea-level potential temperatures using the dry-adiabatic lapse rate of 1 °C per hundred meters. To minimize the effect of highly uncertain extremes, a minimum mixing height of 400 m and maximum of 1200 m were imposed. In future studies, a more sophisticated algorithm may allow more relaxed constraints and more effective use of mixing heights in semi-empirical models.

Because ozone is a photochemical pollutant, a variable representing the total of the hourly k_1 (rate constant for photolysis of NO_2) values by day of the year was included. These calculations assumed clear skies and reflect the changing angle of the sun during the ozone season. We explored using cloud-cover data and the total k_1 values to substitute for the lack of directly measured solar radiation measurements. Although some promising pictures emerged, we did not find a satisfactory substitute for direct data on solar radiation.

Table 2. Meteorological variables used in analyses

Type of data	Summary form (by site)	Units
Data for surface sites		
A. Temperature	Avg. of 10:00-16:00 PST obs.	°C
B. Cloud cover	Avg. of 10:00-16:00 PST obs.	Scores of 0 to 8
C. Sea-level Pressure	Avg. of 10:00-16:00 PST obs.	mb
D. Wind speed	Avg. of 10:00-16:00 PST obs.	ms^{-1}
E. North wind, u	Avg. of 10:00-16:00 PST obs.	ms^{-1}
F. East wind, v	Avg. of 10:00-16:00 PST obs.	ms^{-1}
G. Temperature^2	Square of A.	
H. (Wind speed)$^{-1}$	Inverse of D.	
I. Pressure gradient	Differences in C for selected pairs of sites	mb
J. Temperature gradient	Differences in A for selected pairs of sites	°C
Upper air data		
K. 850 mb temperature	4 p.m. PST at Oakland	°C
Other data		
L. Mixing height	At the first hour of the ozone maximum in the southern subregion	m
M. Total k_1	Daily total of hourly photolysis rates for NO_2 (calculated for clear skies)	

3 Methodology

This section discusses the methods used to construct the two-stage model by first clustering days and then calibrating a model within each cluster to relate ozone to selected meteorological conditions.

3.1 Defining clusters

The first stage in developing the model was to define clusters of similar days. This step used data for 1988 – 1996. The clusters were based on pressure and temperature gradients that represent the synoptic-scale and meso-scale forcing functions that drive the speed and direction of winds. The residence time of an air mass over emission sources and the trajectory of that air mass during the day are determined by the speed and direction of the winds. Accordingly, the days in a cluster tend by to be similar with respect to source-receptor relationships. The limited purpose of the clusters in this study departs from the more extensive role given to clusters in the method developed by Davis, *et al.*

Because the gradients used to cluster the days are correlated and not independent, principal component scores based on the correlation matrix were used as the input to the cluster analysis. The clusters were determined by Ward's criterion with a k^{th} nearest neighbor parameter of 10 to reduce the influence of outliers; a small number of days were excluded from the analysis and not assigned to a cluster because they were too isolated from the rest of the days.

Ward's criterion was used to determine clusters for two reasons. First, it has a strong tendency to produce clusters of similar size, which reduces the incidence of clusters with too few members to be useful. Second, it tends to yield spherical clusters in the coordinate space, which is consistent with the goal of obtaining clusters of days with similar source-receptor relationships. The clustering criteria used here also differ from those used by Davis, *et al.*

The clustering process was stopped at 12 clusters somewhat arbitrarily. Table 3 shows how surface wind speeds and directions reflect the influence of the forcing functions by cluster. The three locations in Table 3 are the Alameda Naval Air Station (WMO 745060) in the center of the Bay Area, Travis Air Force Base (WMO 745160), which is inland in the NE air flow corridor, and San Jose Airport (WMO 724945), which is inland in the SE air flow corridor.

The clusters alone accounted for more than 40 percent of the variability in basinwide daily-maximum ozone concentrations during the ozone seasons of 1993-1995, which were selected because they preceded the introduction of California's reformulated gasoline and because they contained a wide spectrum of ozone concentrations. Although the clusters are associated to some degree with different parts of the ozone season, seasonality alone does not explain the variation in ozone that is accounted for by the clusters. When the full season is broken into 12 sequential time periods to represent seasonality, the periods account for only 13 percent of the variation in the ozone data. The clusters are clearly connected with significant differences in daily meteorology that affect ozone concentrations apart from simple seasonality.

Transactions on Ecology and the Environment vol 29 © 1999 WIT Press, www.witpress.com, ISSN 1743-3541

3.2 Models within clusters

The clusters were not designed to account for all of the variability in ozone that is attributable to meteorology. Therefore, an additive model relating meteorology and ozone (and including an indicator for weekend days) was prepared for each cluster in two steps. First, step-wise regression was used to find important meteorological variables for each cluster. But, variable selection methods, such as step-wise regression, frequently over-fit the data and produce models that fail when applied outside the set of training data. To minimize over-fitting, the second step identified a parsimonious model limited to a few explanatory variables.

Table 3. Relationship of north (u) and east (v) wind vectors to clusters defined by principal component scores for six pressure and temperature gradients.

	Mean pressure gradient			Mean temperature gradient		
Cluster	North	East	South	N.E.	S.W.	East
1	0.43	1.02	-2.36	11.0	6.2	11.1
2	2.02	-1.93	-2.24	4.1	2.6	4.1
3	-4.43	-5.53	-4.80	10.2	5.4	10.6
4	2.05	-1.64	-2.94	7.8	4.3	8.0
5	-1.04	-1.74	-3.87	13.0	6.1	13.1
6	-2.11	-2.33	-3.49	9.2	4.5	9.5
7	-0.73	-5.25	-4.17	7.1	3.5	7.0
8	-4.98	-5.86	-4.09	6.2	3.1	6.0
9	4.17	3.61	-0.58	1.6	1.7	1.3
10	-2.00	-3.45	-1.22	1.2	1.4	1.9
11	3.39	2.26	-1.31	6.6	3.8	6.2
12	-3.63	-5.67	-2.83	3.1	1.7	3.2

	Mean wind speed along north and east vectors					
	Travis AFB		San Jose		Alameda NAS	
Cluster	North	East	North	East	North	East
1	-1.04	-2.53	1.82	-1.10	0.58	-2.95
2	0.31	-1.58	2.22	-1.51	-0.06	-3.17
3	-4.33	-7.38	2.75	-1.71	-1.30	-4.66
4	-1.52	-3.16	1.92	-1.26	-0.09	-2.96
5	-3.64	-5.64	2.33	-1.48	-0.37	-3.74
6	-3.36	-5.33	2.13	-1.42	-0.06	-3.67
7	-3.79	-6.11	2.85	-1.52	-1.20	-4.10
8	-4.40	-7.24	2.38	-1.43	-0.90	-4.72
9	4.01	1.18	1.62	-1.21	0.89	-2.03
10	-1.79	-2.94	0.07	-1.10	-0.87	-2.48
11	1.86	0.03	1.67	-0.97	0.38	-2.11
12	-3.04	-5.01	1.67	-1.61	-0.91	-3.87

Transactions on Ecology and the Environment vol 29 © 1999 WIT Press, www.witpress.com, ISSN 1743-3541

4 Results and Discussion

4.1 Stage 1 – clusters

The mean wind vectors by cluster shown in Table 3 indicate that the clustering procedure largely achieved its objective. The gradients for pressure and temperature generated the expected differences in wind speed and direction. The number of clusters may be slightly larger than needed. For example, clusters 3 and 4 have very similar pressure gradients and similar, though smaller, temperature gradients. Perhaps, these two clusters should be joined and a single within-cluster model used to explain the varying ozone concentrations. For future studies, a more extensive set of gradients (summarized by principal components) might generate even more definitive clusters in stage 1.

4.2 Stage 2 – models within clusters

Table 4 shows the final model for each of the clusters. Models for clusters 7, 8 and 12 have R^2 values below 0.50 for the training data. The maximum ozone concentrations in these three clusters were less than 90 ppb during the 1993-1995 ozone seasons. For the remaining clusters, the smallest R^2 value was 0.66. Though parsimonious models were selected, further safeguards against over-fitting may be advisable.

Table 4. Models for meteorologically standardized ozone within Clusters, based on the training data for 1993 - 1995

Cluster	N	Model	R^2
1	15	= -314.9 + 19.7 ST(LLL) – 0.212 ST^2 (LLL)	0.85
2	85	= 73.7 – 4.5 ST(LLL) + 0.166 ST^2 (LLL) + 5.8 WE	0.81
3	23	= 56.8 – 0.054 MXHT + 1.47 WSN(SFO) + 1.45 ST(SJO) – 6.4 WE	0.81
4	55	= 130.1 – 10.1 ST(LLL) + 0.251 ST^2 (LLL) + 157 WSI(STK) + 2.5 WE	0.72
5	52	= 230.8 – 15.7 ST(TAFB) + 0.371 ST^2(TAFB) + 1.90 WSE(SFO) + 3.5 WE	0.66
6	56	= 11.1 + 2.99 ST(LLL) – 2.65 WS(TAFB) – 2.20 WSN(SJO) + 5.3 WE	0.75
7	68	= 51.6 – 0.016 MXHT + 112 WSI(TAFB) + 8.3 WE	0.25
8	29	= -33.3 + 3.47 ST(TAFB) – 0.247 WSN(SFO) + 9.4 WE	0.26
9	29	= -4.3 + 2.65 ST(FUNS) – 1.31 WS(TAFB) + 1.77 T850 + 6.3 WE	0.66
10	29	= 280.3 – 2.24 CC(SAC) – 23.4 ST(TAFB) + 0.611 ST^2 (TAFB)	0.67
11	33	= -5.6 + 0.111 ST^2 (LLL) + 6.3 WE	0.80
12	44	= 9.2 + 3.57 ST(LLL) – 0.064 ST^2 (TAFB) + 0.77 WSN(SFO) + 5.8 WE	0.42
w/o clusters		= 101.1 – 6.66 ST(LLL) + 0.207 ST^2 (LLL) – 0.55 CC(STK) + 5.8 WE	0.77

CC = cloud cover, MXHT = mixing height, ST = surface temperature, T850 = 850 mb temperature, WE = weekend indicator, WS = wind speed, WSE = east wind speed, WSI = inverse of wind speed, WSN = north wind speed, FUNS = Fort Funston, LLL = Lawrence Live.more Labs, SAC = Sacramento, SFO = San Francisco, SJO = San Jose, STK = Stockton, TAFB = Travis Air Force Base

Transactions on Ecology and the Environment vol 29 © 1999 WIT Press, www.witpress.com, ISSN 1743-3541

4.3 Performance evaluation

The overall ability of the two-stage model to reproduce the observed ozone data is very encouraging. When combined, the models within the clusters have an R^2 of 83 percent. As a measure of model performance, R^2 characterizes the ability of the independent variables (meteorology) to explain or reproduce the values of the dependent variable (ozone). Figure 1 illustrates the strength of the relationship between the observed ozone values and the meteorologically standardized ozone values produced by the two-stage model.

If clusters are ignored, the best fitting model (equation w/o clusters shown in Table 4) achieves a 77 percent R^2. Figure 2 shows the relationship between the observed and the meteorologically standardized ozone when clusters are not used.

The difference of 6 percent between the R^2 values for the models with and without clustering represents a significant reduction in residual variation around the 1-to-1 line. Presuming a corresponding increase in resolution, the two-stage model may make air quality trends more informative when adjusted for meteorology. The accuracy of the meteorologically standardized ozone values produced by a model is important because they provide the reference point for adjusting air quality trends to compensate for meteorological variability.

The performance of the two-stage model compares quite favorably with other efforts to relate ozone and meteorology. For example, Zeldin, *et al.*[7], used multiple regression models with meteorological variables to explain daily maximum ozone in sub-regions of the South Coast Air Basin. Used data from June-September for two years, they achieved R^2 values between 50 and 75 percent. Other studies have generally reported R^2 values up to 75 percent. Differing seasons and other factors may complicate the comparisons, but the performance of our two-stage model compares very well, nevertheless.

5 Conclusions

The two-stage model developed in this study compares very favorably with the most effective tools for relating daily ozone to daily meteorological variables. Because the model relies on routine meteorological data available in many locations, the method may be prove useful in other areas besides the San Francisco Bay Area.

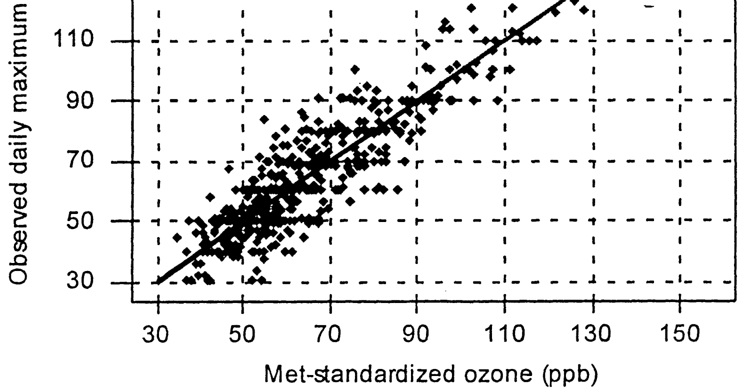
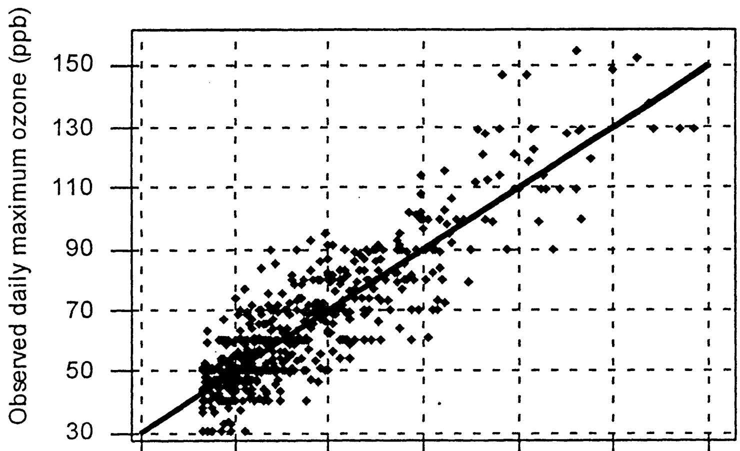


Figure 1. Relationship between basinwide daily maximum ozone in the San Francisco Bay Area and daily met-standardized ozone produced by clustering and within-cluster additive models (data for 1993-1995).



6 References

1. Davidson, A., *An Objective Ozone Forecast System for July through October in the Los Angeles Basin*, South Coast Air Quality Management District, Diamond Bar, CA, 1974.

2. Zeldin, M., & Meisel, W., *Use of Meteorological Data in Air Quality Trend Analysis*, EPA-450/3-78-024, US Environmental Protection Agency, Research Triangle Park, NC, 1978.

3. Cox, W.M., & Chu, S., Meteorologically adjusted ozone trends in urban areas: a probabilistic approach. *Atmospheric Environment*, Vol. 27B, pp. 425-434, 1993.

4. Rao, S.T., & Zurbenko, I.G., Detecting and tracking changes in ozone air quality, J. Air & Waste Management Assoc., Vol. 44, pp 1089-1092, 1994.

5. Davis, J.M., Eder, B.K., Nychka, D., & Yang, Q., Modeling the effects of meteorology on ozone in Houston using cluster analysis and generalized additive models. *Atmospheric Environment*, Vol. 32, pp. 2505-2520, 1998.

6. California Air Resources Board ADAM air quality database web site: http://www.arb.ca.gov/aqd/aqd.htm

7. Zeldin, M.D., Cassmassi, J.C., & Hoggan, M., Ozone trends in the South Coast Air Basin: An update, *Tropospheric Ozone and the Environment*, Air & Waste Management Assoc., TR-19, pp 760-771, 1991.

Acknowledgements

The author thanks the many colleagues at the California Air Resources Board who's assistance made this work possible, particularly Luis Woodhouse for calculating hourly NO_2 photolysis rates and Elliot Mulberg for providing a procedure for calculating mixing heights.