Atmospheric Environment Vol. 16, No. 1, pp. 135–143, 1982
Printed in Great Britain.

0004–6981/82/010135–09 $03.00/0

OXIDANT PREDICTION BY DISCRIMINANT ANALYSIS IN THE SOUTH COAST AIR BASIN OF CALIFORNIA

GONG-YUH LIN
Department of Geography, California State University, Northridge, CA 91330, U.S.A.

(*Received in final form* 4 *February* 1981)

Abstract—Simple and multiple discriminant models of the first stage episode day for the 30-, 24- and 6-h prediction intervals have been derived for Azusa, Burbank, Los Angeles, Fontana and San Bernardino, and those of the second stage episode day have been developed for Azusa and Fontana. Morning weather variables and the previous day's maximum 1-h average oxidant concentration are used as discriminating variables. A scattergram, Pearson and canonical correlation coefficients, and Wilks' Lambda are presented to show the statistical relationship between weather variables and oxidant concentrations or episode occurrences. The morning sounding profiles taken at Los Angeles International Airport differ markedly between episode and nonepisode days at various air monitoring stations. The previous day's oxidant concentration, temperatures at 950- and 850-mb, height of inversion base, inversion magnitude and inversion breaking temperature correlate significantly high with the oxidant concentration at all stations and are important discriminating variables for predicting the occurrence of episode and nonepisode days. It is found that the models yield approximately 65–88% accuracy for the first stage episode day and 51–80% accuracy for the second stage episode day. In the most polluted area, the multiple discriminant model provides very little incremental prediction power over the simple model using the 850-mb temperature or 24-h persistence variable as a predictor for the first stage episode day, but it provides a larger incremental prediction power for the second stage episode day.

INTRODUCTION

Discriminant analysis is a useful statistical tool for deciding whether to forecast, on the basis of observations of several meteorological variables, which of two or more categories will be attained by some predictand. Such categorical forecasts are common in meteorology for precipitation (yes or no, snow or rain), temperature (below freezing, above 86° F or 30° C) and air quality (oxidant level equals or exceeds 0.20 ppm), the last being the application discussed in this report. Specifically, procedures are developed for using morning observations and those of the previous day, in deciding whether to declare first or second stage smog alerts. These are, respectively, when maximum 1-h average oxidant concentrations equal or surpass 0.20 and 0.35 ppm.

The application is for five stations (Los Angeles, Burbank, Azusa, Fontana and San Bernardino) of the South Coast Air Quality Management District, encompassing the greater Los Angeles Air Basin (Fig. 1). Predictors include results from the morning (0600 PST) radiosonde ascent at Los Angeles International Airport (LAX), the previous day's maximum 1-h oxidant concentration at each station, and the sea level pressure gradient across the basin (Table 1).

Approaches for developing statistical models for air quality prediction in Los Angeles have been summarized by Myrabo *et al.* (1977). Earlier, McCutchan and Schroeder (1973) found significant differences in oxidant concentrations in the San Bernardino mountains under different weather patterns, classified by discriminant analysis, but derived no oxidant prediction formulae. Davidson (1974) offered a graphic method for predicting occurrence of summer days with instantaneous peak ozone levels of 0.35 ppm or greater, a "school and health smog warning" criterion abandoned in 1976. Zeldin and Cassmassi (1979) developed a pollution decision tree for one station (Upland) based on automatic pattern recognition, a technique similar to discriminant analysis.

In 1974, the California Air Resources Board (ARB) defined first, second and third stage smog alerts, under which preventive and abatement procedures of increasing degree are required as the maximum 1-h average oxidant concentrations were forecast to equal or exceed 0.20, 0.35 and 0.50 ppm, respectively. After much bickering, these criteria were adopted in 1976 by the South Coast Air Quality Management District (SCAQMD), successor to the Southern California Air Pollution Control District, itself a forced merger of four county smog districts, including the pioneer Los Angeles County Air Pollution Control District. Two years later, in June 1978, the SCAQMD followed the lead of state and federal agencies and changed its method of measuring oxidant concentration from the colorimetric potassium iodide (KI) method to one using ultraviolet photometry (UV). Despite much effort, no consistent correspondence has been found between the KI and UV methods; the UV/KI ratio varies greatly for different stations and months (ARB, 1978). Therefore, only two summers of UV measurements (June–September, 1978 and 1979) are used for analysis.

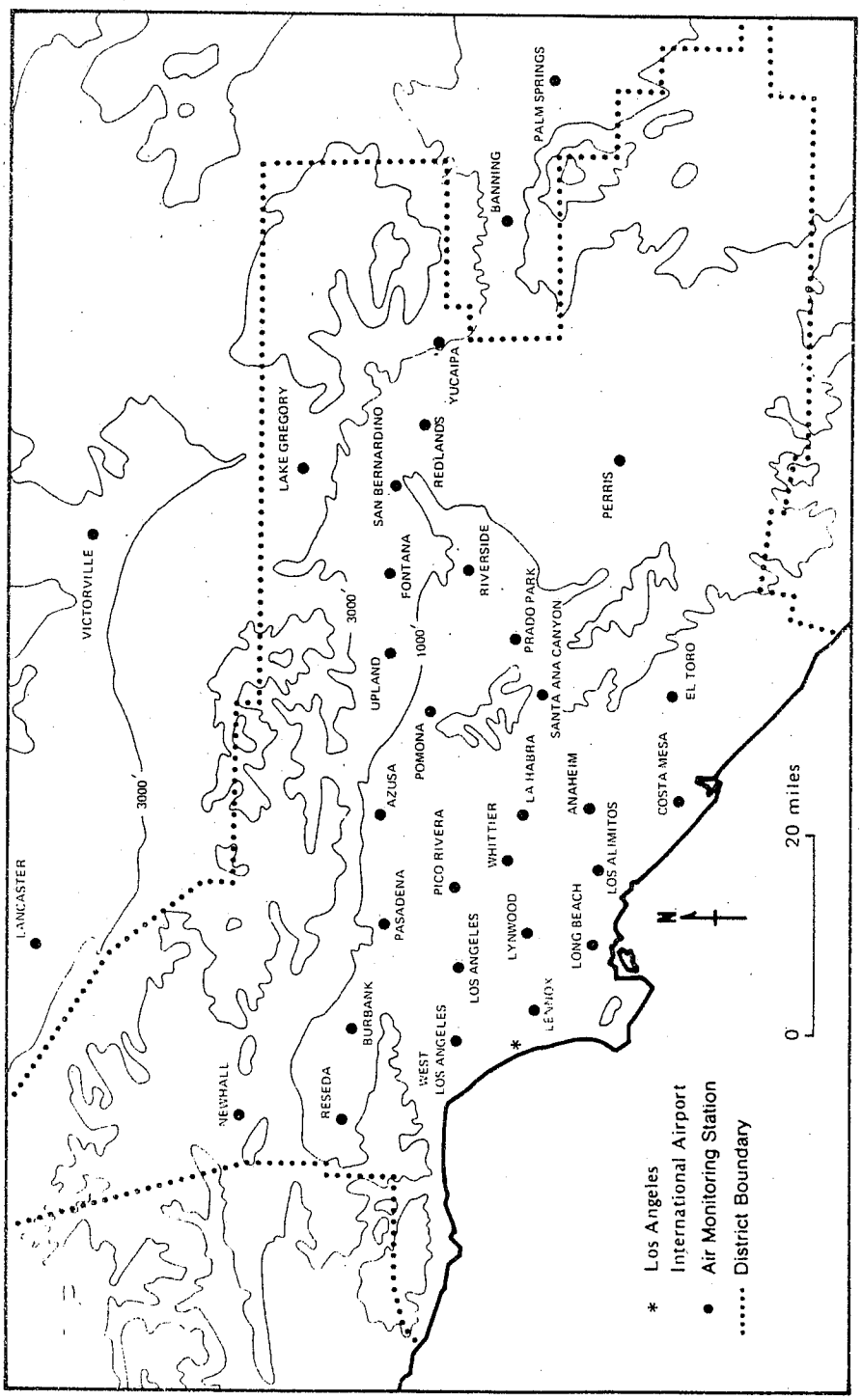

Fig. 1. Locations of air monitoring stations in the South Coast Air Basin of California.

Table 1. Variable list

Codes	Variables
POXT	The previous day's maximum 1-h average oxidant concentration (ppm)
TSFC	Surface temperature (C) at LAX
T950	950-mb temperature (C) at LAX
T850	850-mb temperature (C) at LAX
TINB	Inversion base temperature (C) at LAX
INMG	Inversion magnitude (C) at LAX
TBRK	Inversion breaking temperature (C) at LAX
TD	Surface dew point temperature (C) at LAX
HINB	Inversion base height (in hundreds of meters) at LAX
HINT	Inversion top height (in hundreds of meters) at LAX
PD	0700 PST pressure difference (mb) between LAX and Lancaster William J. Fox Airport

TSFC through HINT are based on 0600 PST radiosonde observations at LAX.

METHOD

Mathematical treatments of discriminant analysis and classification are discussed in detail by several multivariate statistical textbooks (Tatsuoka, 1971; Cooley and Lohnes, 1971) and computer programs are available for the laborious calculations. Discriminant analysis is a statistical technique to determine functions as linear combinations of a set of variables in order to achieve not only the maximum group differentiation but also the minimum probability of error in assigning cases to predetermined groups. A standardized discriminant function is of the form

$$Y = V_1X_1 + V_2X_2 + V_3X_3 + \ldots + V_pX_p,$$

where Y is the discriminant score, X's are the discriminating (predictor) variables and V's are the discriminant coefficients or weights. Both discriminant scores and discriminating variables are expressed in standard values, i.e. zero mean and unit variance. Therefore, a discriminant score is the number of standard deviations from the grand mean (zero score) of all observations of the discriminant function. The discriminant coefficients indicate the relative importance of each variable to the function.

The aim of discriminant analysis is to maximize the ratio of between-groups to within-groups sums of squares of discriminant scores, or F-ratio of variance estimates. When raw scores of discriminating variables are applied, a constant is added to the discriminant function for the adjustment of the grand mean value. In this case, the discriminant weight does not reflect the contribution of discriminating variables to the function because of the differences in units of raw scores.

The number of discriminant functions is one less than the number of groups but no more than the number of discriminating variables. A single discriminant function separates two groups.

Wilks' Lambda, the ratio of the within-groups to the total sums of squares, is an inverse measure of the F-ratio and can be transformed into a chi-square value for testing the significance of a discriminant function (Tatsuoka, 1971).

In discriminant analysis, centroids (group means on the discriminant function) are frequently used as criteria for classification. The midpoint of the distance between centroids serves as a cutoff point, so that cases are assigned to the group to whose centroid their discriminant scores are closer.

Classification can also be achieved by applying raw scores on discriminating variables to a series of classification functions, one for each group, derived from the pooled within-group covariance matrix and centroids for discriminating variables. The classification function is of the form

$$C = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n,$$

where C is the classification score for a given group, a is a constant, b's are the classification coefficients and X's are the discriminating variables in raw scores. The discriminant function is located as the vector which provides the best F-ratio dividing groups while the classification functions, equivalent to multiple regression equations for each group, are the best predictors of scores on the discriminant function for the respective groups (Johnston, 1978).

The following discriminant models of various prediction intervals are derived for each of the five stations:

(1) 30-h prediction length
  (A) multiple weather model
  (B) simple weather model;
(2) 24-h prediction length
  (A) simple persistence model
  (B) 30-h weather and 24-h persistence combination model;
(3) 6-h prediction length
  (A) multiple weather model
  (B) simple weather model
  (C) 6-h weather 24-h persistence combination model.

In general, the 1-h average oxidant concentration peaks at noon or afternoon in the South Coast Air Basin. Therefore, the weather variables observed at 0600 and 0700 h provide approximately 6-h forecast in advance for an episode occurrence on the same day and 30-h prediction length for an episode occurrence on the following day. Since the noon observations of weather variables at LAX are not available for the study period, the morning weather variables are combined with the maximum 1-h average oxidant concentration to derive the multiple discriminant model of an episode occurrence on the following day. The model thus provides a 24-h prediction length. The 850-mb temperature is employed as a single predictor for the simple weather model for the stations other than Los Angeles, where the 950-mb temperature is used as a predictor because it correlates most strongly, among all weather variables, with the oxidant concentration.

The subprogram DISCRIMINANT of an SPSS library program (Nie *et al.*, 1971) is employed to carry out the discriminant analysis of oxidant episode occurrences at the five stations. Options 5, 6, 11 and 12 are specified to obtain classification result tables, discriminant scores and classification informations for all cases, unstandardized discriminant function coefficients, and classification functions, respectively. The other options do not provide informations needed for this report. The forward stepwise screen technique is employed to select variables which contributes a partial F-ratio equalling or surpassing one (default option) as predictors in the multiple discriminate model. The default option, listwise deletion, is used to treat missing data. Cases with missing values are automatically eliminated from all calculations. Thus, all means, standard deviations and correlations are based on the same universe of data. However, the number of cases included in different models varies slightly.

DISCUSSION

Figure 2 shows the mean morning vertical temperature profiles from the ground surface to 850-mb level at Los Angeles International Airport on episode and nonepisode days at Azusa. It can be seen that episode days tend to be associated with higher temperatures in the upper levels. At Azusa, Burbank and Los Angeles, the first stage episode day tends to occur when the 850-mb temperature reaches 15° C or higher (Fig. 3). The corresponding critical temperature for the episode day is 10° C for Fontana and 12° C for San Bernardino.

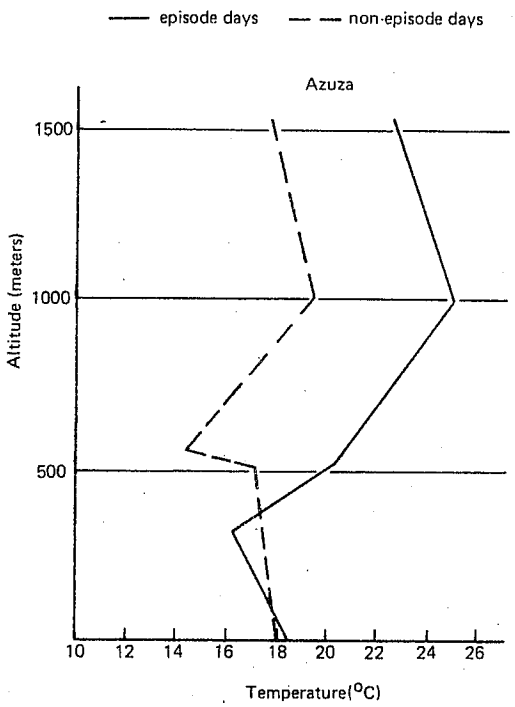

Fig. 2. The mean morning (0600 PST) vertical temperature profiles at Los Angeles International Airport on the episode and nonepisode days at Azusa.

The second stage episode day tends to occur with the 850-mb temperature reaching approximately 20° C or higher for all stations. On the ground surface, the temperature difference between episode and nonepisode days is small. It appears that the episode day occurs with a higher inversion magnitude and with a lower height of inversion base. It also shows that episode days are associated with higher dew point temperatures and the previous day's oxidant concentrations (Table 2). Smaller pressure differences between the coast and desert or even negative values, an indication of weakening onshore flow, favor the occurrence of episode days.

The relationship between oxidant concentrations and weather variables shows geographical variations (Table 3). The 850-mb temperature has the highest correlation coefficients with oxidant concentrations at the stations other than Los Angeles. Correlation coefficients between the dew point temperature and oxidant concentrations vary from 0.39 at San Bernardino to only 0.02 at Los Angeles whereas the 950-mb temperature shows a moderate correlation of 0.61 at Los Angeles but almost no correlation at San Bernardino. Except for Los Angeles, pressure differences show very small correlations with oxidant concentrations. Surface temperature and height of inversion top show no correlations with oxidant concentrations at most stations. Correlation coefficients with oxidant concentrations decline slightly for

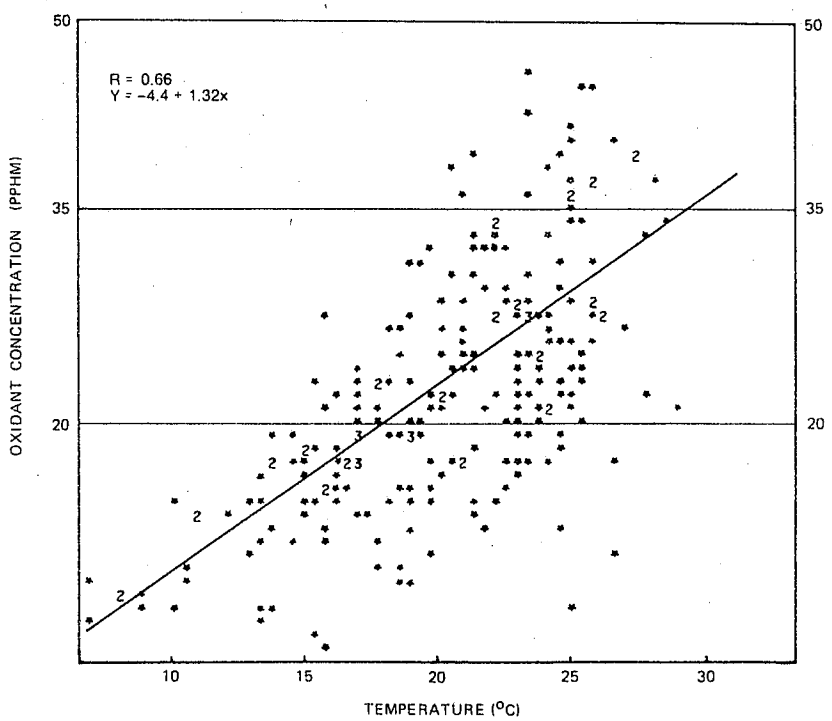

Fig. 3. Scattergram showing the relationship between the oxidant concentration at Azusa and the 850-mb temperature at Los Angeles International Airport.

Table 2. Mean values of the previous day's maximum one-hour average oxidant concentration and morning weather variables for nonepisode and episode days

	Azusa		Burbank		Fontana		San Bernardino	
Variables	Nonepisode	Episode	Nonepisode	Episode	Nonepisode	Episode	Nonepisode	Episode
POXT (ppm)	0.16	0.27	0.14	0.23	0.16	0.26	0.16	0.23
TSFC (°C)	18.06	18.37	18.01	18.78	18.22	18.26	18.08	18.35
T950 (°C)	16.78	20.44	17.15	21.74	17.30	19.82	17.69	20.16
T850 (°C)	17.18	22.63	18.74	23.85	17.12	22.17	17.92	22.88
TINB (°C)	14.36	16.30	14.77	17.12	14.42	16.09	14.76	16.14
INMG (°C)	5.16	8.66	6.24	9.07	5.29	8.27	5.64	8.80
TBRK (°C)	29.13	34.52	30.72	35.68	28.99	34.11	29.71	34.89
TD (°C)	13.98	15.74	14.65	16.25	13.80	15.68	14.18	15.82
HINB (m)	558.16	323.39	510.07	245.23	563.22	341.78	513.34	329.81
HINT (m)	1023.05	1024.42	1056.52	972.38	1012.47	1030.17	1008.41	1043.74
PD (mb)	1.07	0.56	1.17	−0.02	0.74	0.79	0.81	0.71

Table 3. Pearson correlation coefficients between today's and yesterday's oxidant concentrations and between today's oxidant concentrations and weather variables observed today (upper values) and yesterday (lower values)

	POXT	TSFC	T950	T850	TINB	HINB	HINT	INMG	TBRK	TD	PD
Azusa		0.13	0.43	0.66	0.37	−0.47	−0.11	0.50	0.47	0.30	−0.60
	0.65	0.11	0.23	0.49	0.25	−0.31	−0.04	0.35	0.38	0.20	−0.16
Burbank		0.20	0.50	0.66	0.38	−0.43	−0.07	0.50	0.48	0.31	−0.30
	0.73	0.14	0.30	0.54	0.30	−0.37	−0.07	0.38	0.46	0.23	−0.25
Fontana		0.05	0.30	0.65	0.27	−0.39	−0.02	0.50	0.50	0.34	−0.08
	0.63	0.03	0.20	0.48	0.19	−0.27	−0.02	0.39	0.37	0.22	−0.04
Los		0.33	0.61	0.58	0.52	−0.52	−0.24	0.36	0.35	0.02	−0.50
Angeles	0.60	0.26	0.34	0.42	0.39	−0.42	−0.22	0.23	0.31	0.12	−0.45
San		0.02	0.19	0.62	0.20	−0.32	0.05	0.50	0.47	0.39	0.01
Bernardino	0.63	0.00	0.10	0.46	0.13	−0.23	0.09	0.34	0.37	0.29	0.05

the 30-h weather variables as compared with the 6-h weather variables. There are high one-day lag autocorrelations for oxidant concentrations at all stations.

The power of discriminating variables for group differentiation can be judged from the weights associated with individual variables in the standardized discriminant function. Table 4 shows the standardized discriminant function coefficients for the 6-h weather and 24-h persistence combination model. The discriminating power of individual predictors shows distinct spatial variations. At the Azusa, Fontana and San Bernardino stations, the 850-mb temperature is the most powerful discriminating variable whereas the 950-mb temperature and inversion base temperature contributes the most to the group differences on the discriminant functions for Los Angeles and Burbank, respectively.

MODELS

Classification of cases into predetermined groups can be achieved from group centroids and classification functions. For instance, at Azusa the group centroids on the discriminant function for the 6-h weather and 24-h persistence combination model are 0.57 for the episode day and -0.81 for the nonepisode day. The mid-point on the discriminant function is thus -0.12. Cases with discriminant scores greater than -0.12 would be predicted as episode days, and those win scores less than -0.12 as nonepisode days. The discriminant score of each day can be derived either from the standardized or unstandardized discriminant function. An alternative method of classification is through the use of classification functions. The classification function combining the 6-h weather and 24-h persistence as predictors for ~~non-episode~~ nonepisode day at Azusa is

$$C_0 = -41.490 - 7.838(\text{POXT}) + 4.179(\text{TSFC}) + 0.508(\text{T850})$$

and for the episode day is

$$C_1 = -46.289 + 6.411(\text{POXT}) + 3.893(\text{TSFC}) + 0.859(\text{T850}).$$

Raw scores of the discriminating variables can be applied to these functions to obtain the classification scores, C_0 and C_1. Cases would be assigned to the group with the higher score. For example, the C_0 and C_1 values calculated from the observed scores of discriminating variables on 2 June, 1978 are 33.7 and 30.9, respectively. Since C_0 is greater than C_1, the day is classified as nonepisode day, which agrees with the actual occurrence.

The overall accuracy of classification at Azusa using these classification functions is 82.6%. Approximately 83% of nonepisode days and 82% of episode days are accurately classified. The canonical correlation and Wilk's Lambda are 0.67 and 0.54, respectively. The canonical correlation is the maximum correlation between the linear combination function of independent variables and that of dependent variables. In discriminant analysis, dependent variables are group memberships coded with dummy criteria; the episode day is coded 1 and nonepisode day 0.

The other model of various prediction lengths for Azusa is presented in Table 5. The prediction accuracy ranges from 71% for the 30-h simple weather model to 83% for the 6-h weather and 24-h persistence combination model. The improvement in the prediction power of the multiple weather model over the simple weather model is less than 5% either for the 30- or 6-h prediction interval. This is also reflected by a slight decrease in Wilk's Lambda and increase in canonical correlation for the multiple weather model.

The multiple discriminant models derived for the other four stations vary in accuracy from approx. 75–85% for the 30-h prediction length and 80–88% for the 6-h prediction length (discriminant models for these stations are not presented in this report but will be supplied upon request). At Fontana, San Bernardino and Burbank, the multiple weather model improves the prediction accuracy over the simple weather model by 2–12%. The improvement is greatest at Los Angeles, approximately 18 and 23% for the 30- and 6-h prediction lengths, respectively.

In terms of the 24-h prediction length, no significant improvement in the prediction accuracy of the weather-persistence combination model over the simple persistence model are found for Azusa, Fontana and San Bernardino. The improvement is 6% for Burbank and 9% for Los Angeles.

The incremental prediction power of the 6-h weath-

Table 4. Standardized discriminant function coefficients for the 6-h weather and 24-h persistence combination model of the first stage episode

Azusa		Burbank		Fontana		San Bernardino	
Variables	Coefficients	Variables	Coefficients	Variables	Coefficients	Variables	Coefficients
POXT	0.517	POXT	0.447	POXT	0.525	POXT	0.443
		T950	0.196	TSFC	−0.179	T950	−0.412
		T850	0.384	T950	−0.385	T850	0.507
		TINB	−0.604	T850	0.592	TBRK	0.196
		HINB	−0.330	TBRK	0.198	HINB	−0.307
		PD	−0.356	HINB	−0.329		
TSFC	−0.234	TD	0.267	INMG	−0.225		
T850	0.635						

Table 5. Discriminant models of various prediction lengths for the first stage episode day at Azusa

	30-h prediction length				24-h prediction length			
	Multiple weather model		Simple weather model		Weather-persistence model		Persistence model	
Variables	C_0	C_1	C_0	C_1	C_0	C_1	C_0	C_1
POXT					−9.243	11.535	28.587	47.715
T950	1.343	1.191			0.602	0.406		
T850	−0.290	−0.066	1.093	1.368	−1.061	−0.759		
INMG					−0.981	−1.230		
TBRK	1.379	1.462			2.241	2.337		
HINB	2.251	2.054						
PD					1.407	1.232		
Constant	−35.829	−39.233	−9.619	−15.072	−35.746	−41.327	−2.281	−6.356
Wilks' Lambda		0.75		0.77		0.59		0.67
Canonical correlation		0.50		0.48		0.64		0.58
Accuracy (%)		75.1		71.1		79.1		76.4

	6-h prediction length					
	Multiple weather model		Simple weather model		Weather–persistence model	
Variables	C_0	C_1	C_0	C_1	C_0	C_1
POXT					−7.838	6.411
TSFC	6.128	5.919			4.179	3.893
T850	−0.356	−0.113	1.285	1.709	0.508	0.859
INMG	0.100	1.207				
TD	0.690	0.944				
HINB	−0.556	−0.741				
HINT	1.341	1.450				
PD	1.968	1.787				
Constant	−66.045	−72.430	−10.884	−19.297	−41.490	−46.289
Wilks' Lambda		0.58		0.63		0.54
Canonical correlation		0.65		0.61		0.67
Accuracy (%)		79.0		77.8		82.6

er and 24-h persistence combination model over the 6-h multiple weather model is negligible for all stations. At San Bernardino, the weather–persistence combination model even shows a slight decrease of 1% in the prediction accuracy as compared with the multiple weather model for the 6-h prediction length. Such a small reduction in the prediction power could be attributed to the difference in the number of cases involved in the development of different models.

As the prediction interval decreases, the accuracy of the multiple discriminant model increases approximately 10% for Azusa, Burbank and San Bernardino, while there are no significant increasing accuracies for Los Angeles and Fontana. For instance, at Azusa, the prediction accuracy increases from 75.1% for the 30-h multiple weather model to 82.6% for the 6-h multiple weather model. The corresponding increase is from 75.6 to 79.9% for Fontana. The 24-h persistence model yields almost the same accuracy as the 6-h multiple weather model for Azusa and Fontana, where the oxidant concentration is the highest and most persistent in the entire air basin. The 6-h weather and 24-h persistence combination model shows a slight improvement of 4% in the prediction accuracy over the 30-h weather and 24-h persistence combination model for Azusa.

An effort has been made to derive the persistence model using the oxidant concentration at 0600 h as a single predictor to forecast an episode occurrence on the same day. However, the variation in the oxidant concentration at 0600 h is too small to have an F-ratio sufficiently high (equalling or surpassing one) for the development of the persistence discriminant model for Azusa, Burbank and Los Angeles. The model provides an accuracy of 64% for both Fontana and San Bernardino, approx. 17% lower in accuracy than the multiple discriminant model using the early morning weather variables as predictors.

Models for second stage episode days are also derived for Azusa (Table 6) and Fontana, the stations with a large number of second stage episode days for the meaningful study. At Azusa, the prediction accuracy varies from 54.4% for the 30-h simple weather model to 80.4% for the 6-h weather and 24-h persistence combination model. There are significant increases in the prediction accuracy of the multiple weather model over the simple weather model, from 54.4 to 70.7% for the 30-h prediction length and from 57.7 to 80.4% for the 6-h prediction length. For the 24-h prediction length, the weather–persistence combination model increases the prediction accuracy over the simple persistence model from 59.8 to 80.0%. For

Table 6. Discriminant models of various prediction lengths for the second stage episode day at Azusa

	30-h prediction length				24-h prediction length			
	Multiple weather model		Simple weather model		Weather-persistence model		Persistence model	
Variables	C_0	C_1	C_0	C_1	C_0	C_1	C_0	C_1
POXT					6.093	20.757	28.896	43.686
T950					1.843	1.712		
T850			1.022	1.232				
INMG	0.432	0.685						
TBRK					1.052	1.229		
TD	1.856	2.002						
HINB					2.371	2.582		
HINT					−0.031	−0.199		
PD	−0.617	−1.030			1.775	1.296		
Constant	−15.021	−19.247	−10.068	−14.633	−40.163	−46.333	−3.023	−6.909
Wilks' Lambda		0.88		0.92		0.81		0.86
Canonical correlation		0.34		0.28		0.44		0.37
Accuracy (%)		70.7		54.4		80.0		59.8

	6-h prediction length					
	Multiple weather model		Simple weather model		Weather-persistence model	
Variables	C_0	C_1	C_0	C_1	C_0	C_1
POXT					−35.398	−27.728
T950	1.104	1.278			0.734	0.896
T850			0.309	0.380		
TINB	3.179	2.875			1.336	1.095
INMG	0.676	0.927			0.629	0.826
TD	0.809	1.145			1.852	2.073
HINB	3.399	3.257				
PD	2.464	1.943			2.013	1.516
Constant	−51.329	−56.789	−3.014	−4.689	−29.939	−36.424
Wilks' Lambda		0.78		0.97		0.77
Canonical Correlation		0.47		0.18		0.48
Accuracy (%)		80.4		57.7		80.4

the 6-h prediction length, the addition of persistence to weather variables does not improve the prediction accuracy. The weather-persistence combination models yield approximately 80.0% accuracy for either the 24- or 6-h prediction length.

At Fontana, similar results are obtained except that the accuracies of the various models are slightly lower than those obtained for Azusa. The accuracies vary from 51.1% for the 30-h simple weather model to 78.2% for the 6-h multiple weather model.

SUMMARY

This article describes the statistical relationship between weather variables and oxidant episode occurrences at five air monitoring stations in the South Coast Air Basin of California. A scattergram Pearson and canonical correlations and Wilks' Lambda are obtained to show the strength of statistical relation and prediction power of weather variables for episode occurrences.

It is found that the mean vertical temperature profiles taken at Los Angeles International Airport differ significantly between episode and nonepisode days at various air monitoring stations. The previous day's oxidant concentration, temperatures at 950- and 850-mb, height of inversion base, inversion magnitude and inversion breaking temperature correlate significantly high with oxidant concentrations at all stations and are important discriminating variables for predicting the occurrence of episode or nonepisode days. The first stage episode day tends to occur when the morning 850-mb temperature reaches 10–15° C or higher, depending on the location. The second stage episode day tends to occur when the 850-mb temperature reaches 20° C or higher for all stations.

Discriminant models of the first stage episode day for the 30-, 24- and 6-h prediction intervals have been derived for Azusa, Burbank, Los Angeles, Fontana and San Bernardino, and those of the second stage episode day have been developed for Azusa and Fontana. The models provide approximately 65–88% prediction accuracy for the first stage episode day and 51–80% accuracy for the second stage episode day at different stations. In terms of the first stage episode day, the 30-h multiple weather model improves the prediction power over the simple weather model by approximately 10% or less for the stations other than Los Angeles, where the improvement exceeds 20%. The shortening in the prediction length and the addition of persistence to weather variables as one of predictors increase the prediction power at most 10%. At Azusa and Fontana, where the first stage episode day occurs most fre-

quently in the South Coast Air Basin, the simple 24-h persistence model yields an accuracy as high as the 6-h multiple discriminant model.

The improvement in the prediction power of the multiple weather model over the simple weather model is more significant for the second stage episode day than for the first stage episode day. In contrast, the simple 24-h persistence model for the second stage episode does not work as well as for the first stage episode day because the second stage episode day occurs much less frequently than the first stage episode day.

As with other multivariate statistical models, a discriminant model is limited in that it is applicable only for the station where the data of the model are used. Therefore, different models must be developed for various stations in the basin. In this study, the various models derived for the five stations provide moderate degrees of accuracy for predicting episode occurrences. In general, all models work slightly better for the stations in the Los Angeles Basin, including Los Angeles, Burbank and Azusa, located in or near the area of major pollution emission sources, than for the stations located at the farther end of the receptor area, such as Fontana and San Bernardino.

REFERENCES

California Air Resources Board (1978) *California Air Quality Data* (July–August–September) **10**, 2–3.

Cooley W. W. and Lohnes P. R. (1971) *Multivariate Data Analysis*. Wiley, New York.

Davidson A. (1974) An objective ozone forecast system for July through October in the Los Angeles Basin. *Technical Services Report*. Los Angeles Air Pollution Control District. The report was partially reproduced in survey of statistical models for oxidant air quality prediction. *Adv. Envir. Sci. Technol.* **7**, 402–403.

Johnston R. J. (1978) *Multivariate Statistical Analysis in Geography*, p. 243. Longman, New York.

McMutchan M. H. and Schroeder M. J. (1973) Classification of meteorological patterns in Southern California by discriminant analysis. *J. appl. Met.* **12**, 571–577.

Myrabo L. N. *et al.* (1977) Survey of statistical models for oxidant air quality prediction. *Adv. Envir. Sci. Technol.* **7**, 391–422.

Nie N. H. *et al.* (1971) *Statistical Package for The Social Sciences (SPSS)*, pp. 434–467. McGraw-Hill, New York.

Tatsuoka M. M. (1971) *Multivariate Analysis*. Wiley, New York.

Zeldin M. D. and Cassmassi J. C. (1979) *Development of Improved Methods for Predicting Air Quality Levels in The South Coast Air Basin: Final Report*. Technology Service Corporation. Santa Monica, California. A revised version of the report has been accepted for publication in *J. appl. Met.* (1981).