Journal of Official Statistics
Vol. 5. No. 4. 1989. pp. 375–389
© Statistics Sweden

A Comparison of Two Approaches to Classification of Air Pollution Data

A. Narayanan[1] and Thomas W. Sager[2]

Abstract: Weather and emissions are the primary determinants of air pollution. The classification of days into categories in terms of their meteorological potential for producing harmful air pollution aids scientists in understanding air pollution and help regulators in controlling it. The Texas Air Control Board (TACB), the state agency whose responsibilities include monitoring and controlling air pollutants in the State of Texas, recently classified a selected set of days into nine well-defined categories (WPC) for this purpose. Guided by a written protocol and the exercise of professional judgment, meteorologists assigned each day to a WPC category on the basis of their examination of the weather chart for the day and independently of the air pollution level. This is a laborious, time-consuming task. The categories are then used in understanding the movement of air and relating the weather pattern classification to formation of elevated levels of ground-level ozone, a significant air pollutant.

The aim of this paper is two-fold: (1) to imitate the labor-intensive judgmental WPC as nearly as possible by a purely automatic statistical classification based on discriminant analysis, and (2) to determine the extent to which either WPC or the statistical classification (called STATCLASS) successfully discriminates high- from low-ozone-potential days. It is found that STATCLASS was able to assign about 60–70% of the days to the same category assigned by WPC – an agreement rate comparable to that of the TACB in cross-validation checks made on WPC. We also found that both schemes were reasonably successful in discriminating high-ozone from low-ozone days but that STATCLASS was more successful than was WPC.

Key words: Discriminant analysis; multiple regression; stepwise discrimination; stepwise regression; selection of variables; weather pattern classification; air pollution data.

1. Introduction

Air pollution is an important problem afflicting modern urban civilization. In many cities, ambient concentrations of air pollutants regularly exceed levels thought to threaten human health. The economic costs of controlling air pollution are a sizable burden for many governmental units.

Ground-level ozone is a major air pollutant. A bluish, irritating gas of pungent odour, ozone is beneficial in the stratosphere, where it neutralizes ultraviolet radiation, but is maleficent at the earth's surface, where it reacts with plant tissue, building materials, lung tissue, etc., and degrades them. The literature on air pollu-

[1] Department of Decision Sciences, Indiana University, Bloomington, IN 47405, U.S.A.
[2] Department of MSIS, University of Texas at Austin, Austin, TX 78712, U.S.A.

tion and its effects, and ozone in particular, is extensive. A few recent articles are Bedi, Horvath, and Drechsler-Parks (1988) on the pulmonary effect of ozone and Bambawale (1986) on crop injury attributable to ozone.

Ozone is a secondary pollutant. It is not emitted directly into the atmosphere but arises as a by-product of complex chemical reactions involving precursor pollutants (such as nitrogen oxides and volatile organic compounds) in the presence of sunlight. Therefore, weather can markedly aggravate or mitigate ozone levels. Among the substantial literature documenting the relationship between various weather variables and ozone are Ludwig and Shelar (1978), Kelly, Ferman, and Wolff (1986), Vukovich and Fishman (1986), and Altshuller (1988). Hot sunny days with stagnant air and low inversion layers promote the formation of ground-level ozone. Storms and precipitation clean the air of ozone.

In order to protect human health, the U.S. Environmental Protection Agency (EPA) has set National Ambient Air Quality Standards (NAAQS) for major pollutants. The NAAQS for ozone is that an area shall not experience an ambient concentration greater than 12 parts per hundred million at any monitoring site more than three days in any three-year period. About 90 urban areas in the U.S. are in violation of this standard. In Texas, there are four areas in violation. States with areas in violation of NAAQS are required to submit State Implementation Plans (SIPs) showing by EPA-approved modelling methodology how they will bring their areas of violation into compliance. The Texas Air Control Board (TACB), the state agency charged with developing the required SIPs for Texas, expects to spend two years, 1990–1991, in the modelling endeavor and then begin to implement SIP controls in 1992. Other states will have parallel endeavors.

The relationship between weather patterns and ozone levels plays a major role in SIP modelling strategies for reducing ozone. In the two major models (Empirical Kinetic Modelling Approach and the SAI Urban Airshed Model) sanctioned by the U.S. EPA for use in achieving NAAQS, required reductions are tuned to the historical meteorological conditions and emissions inventories that prevailed on specific typical days when exceedances of thresholds occurred. Weather is an externality in these models. Reductions in ozone can be effected only indirectly by reducing emission levels of precursor pollutants. Thus, if weather patterns become more favorable to ozone formation, there may be no change in ozone levels in spite of successful controls on anthropogenic sources of precursor pollutants. Therefore, erroneous evaluations of the effectiveness of SIP ozone control policies may result unless ozone-favorable weather patterns can be identified and incorporated into the data analysis. Evidence of declines have been found, for example, in weather-adjusted ozone in southern California (Davidson (1986)). EPA procedures currently do not make provision for weather considerations in evaluating compliance.

Several attempts have been made to adjust air pollution measurements for weather. Most of these attempts do not use advanced statistical methodology. Frequencies and bar charts are often used to discretize weather classes. Heidorn and Yap (1986) examined summer versus winter ozone within similar weather events. In a subsequent paper, Yap, Ning, and Dong (1988) analyzed ozone episodes under various weather classes. Pollack (1986) presented an index based on temperature, wind speed, and cloud cover to gauge high ozone potential. A few studies have included precursor pollutants. Hough and Derwent (1987) disaggregated ozone

among different chemical species of hydrocarbons. Balentine and Carter (1987), to which we refer later, did use advanced methodology: principal components followed by cluster analysis to identify patterns of weather and precursor pollutants that were favorable to ozone formation.

At the TACB, meteorologists have developed a Weather Pattern Classification (WPC) scheme (Zimmermann, Tropp, and Barta (1987)). This labor-intensive scheme classifies each day into one and only one of ten different categories (nine defined categories and one unclassified or miscellaneous group). Discussed more fully in Section 2, the categories were defined as regularly recurring bundles of reasonably distinguishable meteorological characteristics that are thought by meteorologists to be useful covariates in distinguishing high-ozone-potential from low-ozone-potential days. Only weather variables are used to classify days in the WPC scheme – neither ozone nor other pollutants are used, nor even known to the classifier. For example, a day will be classified as category 2 if there is a tropical storm in the Houston area. This classification is made independently of ozone measurements for the day; but storms are known to be associated with reduced ozone.

Some of the important questions associated with the WPC scheme are: Can the classification be routinely applied? Does WPC effectively discriminate days with high ozone from days with low ozone? In other words, can WPC serve as a basis for adjusting ozone for weather? And, can WPC be improved? The difficulty with routine application of WPC is that, although based on objective criteria, the classification of days by WPC requires judgment and evaluation by trained professionals. This is so time-consuming and labor-intensive that only Houston has been classified and then only for the high ozone season (May–October) for certain years. If the classification procedure could be automated, considerable time of scientific personnel could be freed, the research programs based on WPC could be advanced, and modelling efforts for regulation could be more soundly based. But incorporating the scientific insights of human classifiers into an expert system is not yet feasible, given existing constraints on resources. As an alternative, the possibility of classifying days by purely statistical criteria is an attractive option. The Radian Corporation recently performed a study under contract to investigate the feasibility of a statistical approach (Balentine and Carter (1987)). This study produced an alternative classification scheme based on cluster analysis after preliminary variable screening and a principal components analysis on the selected variable set. The purpose of their study was to find "natural" clusters in the data and select precursor pollutant data to see if the resulting clusters nevertheless retained the ability to discriminate high from low ozone.

In our paper, the focus is somewhat different. We have two main objectives: (1) to mimic the labor-intensive judgmental WPC as nearly as possible by a purely automatic statistical classification (called STATCLASS) based on discriminant analysis, and (2) to determine the extent to which either WPC or STATCLASS successfully discriminates high- from low-ozone-potential days. If a computer-based automatic classification scheme can describe weather patterns that usefully distinguish high from low ozone and that agree reasonably well with the judgments of meteorologists, then the remaining Houston data and other sites in Texas and throughout the U.S. can be quickly classified. This may make it feasible to incorporate weather adjustments into scien-

tific input to the administrators who will be evaluating the progress of SIP attainment in the coming years. Section 2 discusses WPC in detail; Section 3 discusses STATCLASS. In Section 4, we compare and evaluate the two classifications. In Section 4, we also offer some observations on improving WPC for predictive/explanatory purposes.

To anticipate the conclusions, we found, in brief, that STATCLASS was able to assign about 60–70% of the days to the same category assigned by WPC – an agreement rate comparable to that of the TACB in reassignment checks made on WPC. We also found that both schemes were reasonably successful in discriminating high-ozone from low-ozone days but that STATCLASS was more successful than was WPC.

1.1. Data

The data for the statistical classification scheme (STATCLASS) of this study were collected in Harris County, Texas. This region includes the Houston metropolis, second only to Los Angeles among U.S. cities in significance for ozone pollution. The data were collected through the combined efforts of the Houston Regional Monitoring (HRM) authority, Texas Air Control Board, and the City of Houston. HRM has performed continuous ambient air quality analysis of the Harris County area since 1981. At present the HRM program consists of several monitoring sites in and around Houston which take continual meteorological and air pollution measurements throughout the day. Currently, weather, air pollution, and WPC data are simultaneously available only for the years 1982 and 1983. Because WPCs have been assigned only for days in the ozone season (May–October), only those six months were analyzed. Thus, daily data for two six-month periods totalling 368 days are available. A few of those days have missing values in one or more variables; a day was omitted from any analysis using a variable for which the day's value was missing.

The variables we used are widely recognized as influential covariates in analyses of ozone (e.g., Altshuller (1988), Ludwig and Shelar (1978), Kelly, Ferman, and Wolff (1986), Pollack (1986), and Balentine and Carter (1987)). In fact, we began with the same 20 meteorological and non-ozone pollution variables as in the Balentine-Carter study. To this list we added precipitation, which was effective in lowering the misclassification rate. And we deleted the area average diurnal temperature range because it was collinear with two of the other variables, being equal to AVGTMAX-AVGTMIN. Table 1 lists the 20 variables we used. Five are non-ozone pollution variables; 15 are meteorological; none involves ozone. The pollution variables are various chemical species and measures of nitrogen oxides, which are chemical precursors of ozone.

2. The WPC Classification Scheme

The TACB meteorologist begins his or her other classification of a particular day with an examination of the daily surface synoptic weather charts for the day. The meteorologist also has access to hourly weather observations from Houston airports, wind direction changes, local sea breeze formation, etc. But the meteorologist has no access to ozone or to any other pollution data. After following the objective classification protocol (shown in Figure 1) and exercising professional judgment where required, the meteorologist will classify that day into one and only one of ten broad categories, shown with their descriptive

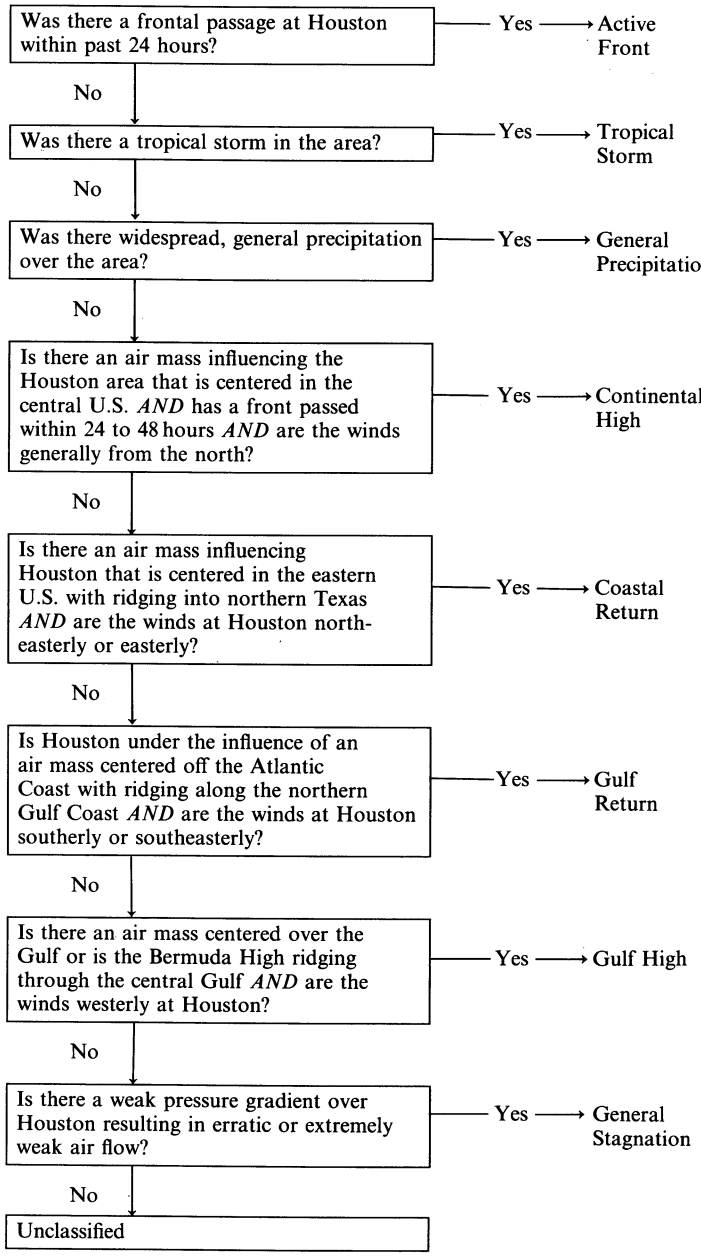

Fig. 1. Protocol for weather pattern classification

labels:

1. Active Front
2. Tropical Depression or Storm
3. General Precipitation
4. Continental High
−5. Coastal Return without Sunshine
+5. Coastal Return with Sunshine
6. Gulf Return
7. Gulf High
8. General Stagnation
9. Unclassified.

Each of the above categories is described in detail by Zimmermann, Tropp, and Barta (1987). For the purpose of our analysis only the category numbers are important. Thus the WPC for a day is one of the above ten numbers. Each of the 368 days of this study has a WPC number available to us. Because category 9 is a group of "leftover" days, which did not fit into any of the other categories and which do not follow a well-defined meteorological pattern, we elected to omit category 9 and its 9 days from our analysis. Thus, the maximum number of observations used in an analysis is 359.

Since the WPC classification procedure intends maximal objectivity, it is required that the meteorologist have no personal bias or preconceived notions of what the weather might be on that day. He or she is trained to be objective and consistent. It is particularly crucial that the meteorologist have no knowledge of the ozone level for the day to be classified, so that he or she assigns the WPC independently of knowledge of the ozone level.

However, the *definition* of the nine WPC categories is not independent of ozone. That is, the weather types shown above were chosen for their presumed ability to discriminate ozone, not simply because they also represent identifiable, recurring *types* of weather. In fact, we shall see that the WPCs do have power to discriminate ozone. Because of the strong dependence of ozone formation on weather, when similar weather patterns recur, similar ozone levels tend to recur as a response.

3. The Statistical STATCLASS Classification Scheme

The basis for the statistical classification of days into meteorologically significant categories is multiple discriminant analysis (Krzanowski (1988)). In discriminant analysis a rule is devised through a function which maps the values of explanatory weather and precursor pollutant variables into the nine categories given above. The objective of this STATCLASS rule is to minimize the disagreement of STATCLASS-assigned categories with WPC-assigned categories among the 359 days. In the classical Fisherian formulation of discriminant analysis, the rule is to assign a day to the closest group in the sense of Mahalanobis distance: $(\mathbf{x} - \bar{\mathbf{x}}_i)'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i)$. Thus, a day with covariate values $\mathbf{x}$ would be assigned to the group $\mathbf{i}$ which minimized the Mahalanobis distance, where $\bar{\mathbf{x}}_i$ is the vector of mean covariate values for group $\mathbf{i}$ and $\mathbf{S}$ is the pooled within-group covariance matrix. This classification rule is equivalent to selecting the group with maximum posterior membership probability when using multivariate normal likelihood and equal priors on the groups. We used a slightly different version of this rule, with priors proportional to number of WPC days in the groups, after examining the reasonableness of the assumption of similar within-group covariance matrices. Classical discrimination works best when the explanatory variables follow a multivariate normal distribution. In other, nonparametric formulations of discriminant analysis, the category assignment of a

*Table 1. List of variables used**

sr	AVGNXRAT	Average of 06-09 CST site NO_2/NO_x ratios
s	DIFFUSIV	Vertical eddy diffusivity for 14 CST ($m^2.sec^{-1}$)
s	DLSLOOZ	24 hour sea level pressure change (kPa) at 00 UTC
s	MAZ10	Height change (m) of 100 kPa surface from 00 UTC to 12 UTC
s	MAZ85	Height change (m) of 85 kPa surface from 00 UTC to 12 UTC
sr	OPQAVG	Average total opaque sky cover for 07-11 CST
sr	RAIN	Precipitation in inches at Houston International LCD
s	SEAM2	Wind speed component of ARWSAM2 from the southeast
s	ZARSTGAM	LOG (ARSTGAM2) where ARSTGAM2 is inverse of area average resultant wind speed for 06-09 CST
sr	ZARVWSAM	LOG (ARVWSAMD) where ARVWSAMD is area average resultant wind speed for 10–13 CST
sr	ZAVGITNR	LOG (775-AVGITNR) where AVGITNR is area average non-negative daily net radiation (W)
	ZAVGNOXM	LOG (AVGNOXMX) where AVGNOXMX is area maximum 1-hour NO_x concentration (ppb)
r	ZAVGPNO	LOG (AVGPNO + .5) where AVGPNO is area average 06-09 CST NO concentration (ppb)
sr	ZAVGPNOX	LOG (AVGPNOX + .5) where AVGPNOX is area average 06-09 CST NO_x concentration (ppb)
r	ZAVGPNO2	LOG (AVGPNO2 + .5) where AVGPNO2 is area average 06-09 CST NO_2 concentration (ppb)
sr	ZAVGTMAX	LOG (100-AVGTMAX) where AVGTMAX is area average daily maximum temperature (°C)
sr	ZAVGTMIN	LOG (85-AVGTMIN) where AVGTMIN is area average daily minimum temperature (°C)
sr	ZHOZ71	LOG (3100-HOZ71) where HOZ71 is thickness (m) between 100 kPa and 70 kPa levels at 00 UTC
r	ZINDEXTO	LOG (65-INDEXTO) where INDEXTO is K index computed from 00 UTC sounding
sr	ZTOZ10	LOG (35-TOZ10) where TOZ10 is temperature (°C) of 100 kPa (1000 mb) level at 00 UTC

*"s" in first column indicates variable was used in STATCLASS.
"r" in second column indicates variable was also used in regression analysis.

day is based on the WPC assignments of days that are nearby in the space of the explanatory variables (nearest neighbor or kernel discriminant analysis). We did not pursue such alternative approaches because they did not seem as promising in preliminary analysis as the classical approach.

The covariates used in the discrimination

are shown in Table 1. The inclusion of precursor pollutant measurements in STATCLASS and their exclusion in WPC may give STATCLASS an advantage. But both WPC and STATCLASS are unbiased in excluding ozone from classification criteria, since each is intended to discriminate high ozone days. As noted in the introduction, the meteorological variables are thought to be among the more significant ones for ozone formation.

In order to better match the assumption of multivariate normality, several of the covariates were transformed so that their univariate histograms appeared more normal (i.e., symmetric and bell-shaped). Each transformation reduced the Kolmogorov-Smirnov distance between the transformed covariate distribution and a normal distribution with same mean and standard deviation to under 0.10. Although this does not *guarantee* multivariate normality, the transformations significantly improved the explanatory power of STATCLASS models and increased the agreement between STATCLASS and WPC. In a few cases, no transformation was applied, even though the covariate distribution was obviously non-normal. For example, precipitation (RAIN) has a very large probability mass at zero, and opaque sky cover (OPQAVG) is U-shaped.

4. Analysis of STATCLASS and WPC

Our primary objective is to match the WPC classifications as nearly as possible with STATCLASS. But we are also mindful of the likelihood of information redundancy among our covariates. Therefore, we first ran a stepwise discriminant analysis (PROC STEPDISC in SAS), at each step of which variables may enter or be removed if the significance of their partial F-test is less than or greater than 0.15, respectively. The resulting model retained 16 of the 20 variables and correctly matched 220 of the 342 possible days on WPC. See Table 2; entries on the main diagonal tally correct matches. Note that the posterior assignment proportions closely match the prior (empirical) WPC proportions.

Interestingly, three of the four variables eliminated in the STEPDISC procedure were air pollution variables: ZAVGPNO, ZAVGPNO2, and ZAVGNOXM (the fourth was ZINDEXTO). Because WPC assignments are made without explicit knowledge of any air pollution variables, it is perhaps not surprising that air pollution variables are of relatively little value in discriminating WPC assignments. However, the question is raised as to whether the WPC scheme could be improved in terms of discriminating ozone by explicitly including non-ozone pollutants as covariates. We return to this question at the end of this section.

Some of the remaining covariates may be eliminated without unduly increasing the misclassification rate. For example, if TOZ10 and AVGPNOX (another air pollution variable) are eliminated, the resulting STATCLASS correctly matches 215 of 342 WPC assignments. If MAZ85 and AVGITNR are also removed, the number of correct matches falls to 208 of 342. Interestingly, if all 20 covariates are retained, the number of correct matches is only 215 of 342. If the 15 weather variables are the only covariates, the number of correct matches is 212 of 342. If the 5 air pollution variables are the only covariates, the number of correct matches is only 120 of 342; and a disproportionate 198 days are assigned to category 6. The relatively good performance of the all-weather-variables discrimination in comparison with the all-pollution-variables discrimination further emphasizes the close connection of WPC with meteorology and its independ-

Table 2. Discriminant analysis classification summary

From WPC	Number of observations and percent classified into WPC									
	1	2	3	4	−5	+5	6	7	8	Total
1	15	0	3	0	1	3	3	1	6	32
2	0	3	0	0	0	0	2	0	1	6
3	3	0	16	0	2	0	8	0	2	31
4	0	0	0	22	2	7	0	0	0	31
−5	1	0	1	1	18	5	5	0	4	35
+5	0	0	0	6	3	28	2	2	1	42
6	0	0	1	2	1	5	68	5	8	90
7	0	0	0	0	0	1	3	17	7	28
8	1	0	3	1	1	4	0	4	33	47
Total										
	20	3	24	32	28	53	91	29	62	342
	5.85	0.88	7.02	9.36	8.19	15.50	26.61	8.48	18.13	100.00
Priors										
	32	6	31	31	42	35	90	28	47	342
	0.09	0.02	0.09	0.09	0.12	0.10	0.27	0.08	0.14	1.00

ence of precursor pollutants. We chose to retain the original 16 variables selected by PROC STEPDISC in order to allow greater flexibility in subsequent analyses.

How well does STATCLASS (based on the 16 selected variables) compare with WPC in terms of ability to discriminate ozone? In Table 3, we have given the mean and standard deviation of area peak ozone (ARO3PK) by category for STATCLASS and WPC. One-way analyses of variance of LOGO3PK (the natural log of ARO3PK) on the STATCLASS and WPC categories have R^2 values of 0.41 and 0.31, respect-

Table 3. Summary statistics of ozone by STATCLASS and WPC
Variable: ARO3PK

Class	STATCLASS			WPC		
	N	Mean	Standard deviation	N	Mean	Standard deviation
1	20	11.700	6.25	32	12.906	7.52
2	3	12.667	6.65	61	1.667	4.76
3	24	7.833	3.44	31	7.839	4.04
4	32	11.500	4.66	31	11.645	4.81
−5	28	11.464	3.47	35	11.742	2.91
+5	53	14.641	4.48	42	14.642	5.25
6	91	7.219	2.63	90	7.967	3.39
7	29	11.068	4.63	28	12.071	6.34
8	62	15.830	6.06	47	15.255	4.76

ively. Area peak ozone is the maximum of all the hourly ozone measurements at all of the monitoring sites in the Houston area. Taking the log of ARO3PK helps make it more nearly normal.

From Table 3, we see that both classifications achieve a reasonable spread of the mean ozone values across the categories. Six of the nine standard deviations are smaller for STATCLASS. The mean levels tend to be elevated when expected on meteorological grounds. For example, categories +5 and 8 have the highest mean ozone levels in each classification. Both are meteorologically favorable to the formation of elevated ozone levels (no recent frontal passage or rain, and plenty of sunshine or stagnant air masses or both). Categories 3 and 6 have the lowest mean ozone levels in each classification. With clean air blowing in from the Gulf of Mexico or cleansing general precipitation, both categories favor reduced ozone levels. In the analysis of variance, WPC explains 31% of the variation of LOGO3PK, whereas STATCLASS explains 41%. By comparison, the discriminant classifications based on all 20 variables, on only the weather variables, and on only the air pollution variables explain 39%, 37%, and 23%, respectively, of the variation of LOGO3PK.

The TACB recently tested its WPC assignment procedure by cross-checking the assignments for 1981, a year not available for our study because of the lack of matching covariate data. A meteorologist independently reclassified each day, using the same WPC definitions but a somewhat different protocol. We then achieved 60% agreement between the original WPCs and reclassified WPCs. Since STATCLASS matched 64% of the original WPCs (albeit for 1982 and 1983, which were not cross-checked), we judge its performance to be at least acceptable.

However, it must be remembered that STATCLASS has been *tuned* to the WPCs assigned in 1982 and 1983. The discrimination attempts to match those WPC assignments as closely as possible. STATCLASS would not be expected to do as well in attempting to match the WPC assignments to days for which the WPC assignments either had not yet been made or were unknown. To test how well STATCLASS would perform in such circumstances, we conducted a cross-validation. The 359 days from 1982 and 1983 were split into two nearly equal groups by randomly assigning one of each pair of consecutive days to a training set group and the other day in the pair to a test group. The STATCLASS discrimination rule was developed on the training set in the usual way, with the 16 covariates previously chosen and with access to the WPC assignment for the days in the training set group. Then the STATCLASS rule developed on the training set was applied to the test group *without* knowledge of the WPCs assigned to the test group. After STATCLASS had made its assignments to the test group, the STATCLASS assignments were compared with the WPCs actually assigned to the test group. In the training set group, STATCLASS correctly matched 110 of 169 days. In the test group, as expected, STATCLASS matched fewer days correctly. But 93 of 173 days were still matched correctly in the test group, for a 54% success rate. In addition, the training set classification explained 36% of the variability of LOGO3PK for the days covered by the training set. And the test set classification explained 32% of the variability of LOGO3PK for the days covered by the test set. In view of our previous remarks about the TACB cross-check, we view these results as acceptable.

The question of whether there are systematic differences between the years 1982

and 1983 that affect the discrimination may be answered in a similar manner. The discrimination may be developed in one year as a training set and applied to the other year as a test set. When the discrimination is run on 1982 as a training set (again using the 16 variables previously selected), there are 107 correct matches out of 177 within the 1982 training set year, for a 60% success rate. When the rule developed for 1982 is applied to 1983, there are 66 matches out of 165 days for a 40% success rate. The corresponding results from using 1983 as training set are 117 matches out of 165 for a 71% success rate within the training set year, and 75 of 177 for a 42% rate when 1983 is applied to 1982. In view of the somewhat lower success rate on the years as test sets (40%, 42%) than on the test set as a pairwise random split (54%), there may be some meteorologically significant differences between 1982 and 1983.

Throughout the analysis, we have utilized the pooled within-group covariance matrix in spite of the rejection of a formal test for homogeneous covariance matrices at the 10% level (although the test was accepted at that level for each year separately). In this case we judge the use of the pooled covariance matrix to be conservative. An alternative procedure which uses separate covariance matrices initially occasioned great excitement when it obtained 261 matches out of 342, for a 76% success rate. And even more so for separate-covariance rules developed on each year separately: 1982 had 163 of 177 matches, and 1983 had 155 of 165.[3] However, it soon became apparent that these miraculous rates were the result of overfitting due to loss of degrees of freedom. When the rule developed on 1982 as a training set was applied to 1983, there were only 18 of 165 matches; and for 1983 applied to 1982, only 32 of 177. These results are consistent with the frequently noted phenomenon in discrimination studies that separate-covariance rules often yield higher matching rates with a training set than will pooled-covariance rules, but lower rates with a test set. This illustrates the importance of validation studies. In the present study, there are sufficient data for analysis based on pooled matrices but not on separate matrices. In order to use the approach based on separate covariances, it seems necessary to eliminate more covariates. When this is done, the number of matches in the test set increases, but the number of matches in the training set decreases. For example, suppose that the 7 covariates ZAVGTMIN, OPQAVG, SEAM2, RAIN, ZAVGTMAX, ZARSTGAM, and DLSLOOZ are retained and all others eliminated. Then with 1982 as training set, 101 matches out of 182 are achieved in 1982, but 59 of 166 in 1983. And with 1983 as training set, there are 102 successes of 166 in 1983, but 66 of 182 in 1982. The success rate for these seven covariates on 1982 and 1983 combined is 193/348 = 55%, again based on separate covariances.

Can WPC be improved? It seems likely that at least three lines of work might be profitable. First, we note that if the purpose of WPC is to provide a predictive or explanatory model for ozone, then a multiple regression model is called for. (For example, see Langstaff and Pollack (1985) on variable selection for regression and classification, and Inoue, Hoshi, and Taguri (1986) on predicting nitrous oxide by regression.) One does not expect indicator variables for categories, whether they be WPC or STATCLASS categories, to account

[3] Sometimes the number of possible matches varies when using a different set of explanatory variables. The reason is that only *complete* observations are used in the analysis.

for variation in ozone as well as continuous predictors can. To this end of prediction or explanation, all relevant predictors should be included, not just weather variables. The analysis can provide for separating out and measuring the relative contribution of the weather variables vis-a-vis the others. Second, if the purpose of WPC is to provide readily identifiable and readily interpretable clusters of weather and other variables which have some predictive or explanatory power for ozone, then the issue is whether better clusters can be found. Presumably, the advantage of nominal categories over continuous regression models lies in the greater understandability of the former by the public and by political overseers. Cluster analysis may prove useful here (Balentine and Carter (1987)). Third, a hybrid approach which combines the meteorologist's insight with statistical models may prove most fruitful of all, particularly if the labor of the meteorologist can be significantly reduced.

We shall illustrate the possible improvements resulting from the first line of work (regression). We began with a stepwise regression of LOGO3PK on all 20 predictors in Table 1, using the MAXR procedure.

Table 4. Regression results on predicting ozone

Dependent variable: LOGO3PK

Analysis of variance

Source	DF	Sum of squares	Mean square	F value	Prob > F
Model	13	58.70	4.51	64.4	0.0001
Error	332	21.97	0.07		
Total	345	80.67			

Root MSE = 0.25 R-Square = 0.73

Parameter estimates

Variable	DF	Parameter estimate	Standard error	T for H_0: Parameter = 0	Prob > | T |
INTERCEP	1	3.64	0.23	15.83	0.000
ZAVGPNO	1	−0.07	0.04	−1.75	0.050
ZAVGPNO2	1	0.55	0.12	4.58	0.000
ZAVGPNOX	1	−0.26	0.13	−2.00	0.054
ZAVGITNR	1	−0.08	0.03	−2.66	0.007
ZAVGTMIN	1	0.35	0.07	5.00	0.000
ZAVGTMAX	1	−0.40	0.09	−4.44	0.000
RAIN	1	−0.05	0.03	−1.66	0.144
ZARVWSAM	1	−0.66	0.05	−13.20	0.000
ZTOZ10	1	−0.11	0.07	−1.57	0.095
ZHOZ71	1	0.10	0.05	2.00	0.049
OPQAVG	1	−0.20	0.01	−2.00	0.010
ZINDEXTO	1	−0.11	0.05	−2.20	0.024
AVGNXRAT	1	0.48	0.10	4.88	0.000

For each $k = 1, \ldots, 20$, MAXR produces the "best" k-variable model in the following sense: Having found the "best" $k - 1$ variable model, MAXR adds the variable which produces the greatest increase in R^2 as in ordinary forward stepwise regression. With the k model variables thus preliminarily identified, MAXR then examines the effect on R^2 of each possible pairwise replacement of a model variable with a variable not in the model. MAXR then makes the replacement that produces the greatest gain in R^2. When this process has been completed for each k, the user then chooses one of the 20 models. In making this choice, we were guided by standard criteria such as Mallows' C_p. The final model selected is shown in Table 4.

Standard diagnostic checks support the adequacy of the model, (cf. Cook and Weisberg (1982) for details on these diagnostics.) Plots of the residuals against the predictors are absolutely featureless. The Kolmogorov-Smirnov distance between the distribution of the residuals and the normal distribution with the same mean and variance is 0.028, which is consistent with normality at a significance level better than 0.15. There are no noteworthy outliers or influential points (using Cook's D). Multicollinearity is not a problem (using condition numbers and variance proportions). There is a small amount of first-order autocorrelation (0.226), which is significant by the Durbin-Watson statistic (1.545). Addition of one-day-lagged LOGO3PK as a predictor largely eliminates the autocorrelation, but has little effect on the other coefficients, little effect on the residuals, and little effect on the R-square (0.744). Conceivably, lagged ozone might prove modestly useful in a classification scheme.

We cross-validated the model in two different ways. First, we randomly split each pair of consecutive days into a training set and a test set. We developed the regression model on the training set ($R^2 = 0.77$), then calculated the residuals when the training model was fit to the test set. These "residuals" had a distribution with slightly heavier than normal tails, but very close to zero mean. A "quasi-R^2" (1-SS "residuals"/SS Total) was calculated to be 0.59. Second, the data from 1982 was treated as a training set and the resulting model applied to 1983 as a test set. Then the roles of the years were reversed. The R^2 values were 0.76 and 0.71 and quasi-R^2's 0.56 and 0.71. Again, the "residual" distributions were slightly heavier-tailed than normal, but with means very close to zero.

In Table 4, we note that 4 of the 5 air pollution variables were selected for the regression, in contrast to the stepwise discrimination procedure, which eliminated three of them. We also note that the 13 variables selected explain a relatively high

Table 5. List of abbreviations used in the text

EPA	Environmental Protection Agency
HRM	Houston Regional Monitoring Authority
NAAQS	National Ambient Air Quality Standards
SAI	Systems Applications Inc.
SAS	Statistical Analysis System
SIP	State Implementation Plans
STATCLASS	Statistical Classification/Discrimination Scheme
TACB	Texas Air Control Board
WPC	Weather Pattern Classification

73% of the variation in LOGO3PK. Since WPC explained 31% and STATCLASS explained 41% of the variability of this variable, there is room to improve nominal classifications. We also observe that the regression model weathered cross-validation somewhat better than the discriminant analysis. Differences between 1982 and 1983 in terms of the interaction between weather and precursor emissions may account for this. On the basis of our analyses, we would recommend that a classification scheme oriented toward ozone discrimination include precursor variable considerations. The ozone variation within weather classes shown in Table 3 is too large relative to the between-class variation for precursors not to play an important role.

5. Discussion

We set out with the objective of duplicating by an automatic, statistical means the time-consuming judgmental classification of days into categories based on their meteorological potential for ozone formation. The STATCLASS scheme matches the judgmental WPC about as well as does another judgmental classifier. STATCLASS has somewhat greater ability to discriminate high-ozone from low-ozone potential days than does WPC. Part, but not all, of that advantage results from the use by STATCLASS of precursor pollutants and their exclusion from WPC. In general, weather seems to play a primary role in explaining ozone, with precursor pollutants secondary. This is seen more clearly in the regression model of Section 4. There, the first few predictors of ozone that enter are weather variables. The regression model indicates that weather and precursor variables together can account for a substantial proportion of the variability of ozone. We conclude that both WPC and STATCLASS are useful in categorizing days in terms of ozone potential if simple indices are desired, but that for predictive or explanatory purposes, continuous models can offer substantial gains.

6. Acknowledgement

The Authors would like to thank the assistance of Dr. M.W. Hemphill of Texas Air Control Board for producing us with the data.

7. References

Altshuller, A.P. (1988): Some Characteristics of Ozone Formation in the Urban Plume of St. Louis, Mo. Atmospheric Environment, 22, pp. 499–510.

Balentine, H.W. and Carter, J.C. (1987): Development of Ozone Climatology for Harris County, Texas. Paper presented in 80th Annual Meeting of Air Pollution Control Association, June 1987, New York.

Bambawale, O.M. (1986): Evidence of Ozone Injury to a Crop Plant in India. Atmospheric Environment, 20, pp. 1501–1503.

Bedi, J.F., Horvath, S.M., and Drechsler-Parks, D.M. (1988): Reproducibility of the Pulmonary Function Response of Older Men and Women to a 2-hr Ozone Exposure. Journal of the Air Pollution Control Association, 38, pp. 1016–1019.

Cook, D. and Weisberg, S. (1982): Residuals and Influence in Regression. Chapman and Hall, New York.

Davidson, A. (1986): Comment on "Ten-Year Ozone Trends in California and Texas." Journal of the Air Pollution Control Association, 36, p. 597.

Heidorn, K.C. and Yap, D. (1986): A Synoptic Climatology for Surface Ozone Concentrations in Southern Ontario,

1976–1981. Atmospheric Environment, 20, pp. 695–703.

Hough, A.M. and Derwent, R.G. (1987): Computer Modelling Studies of the Distribution of Photochemical Ozone Production Between Different Hydrocarbons. Atmospheric Environment, 21, pp. 2015–2033.

Inoue, T., Hoshi, M., and Taguri, M. (1986): Regression Analysis of Nitrous Oxide Concentration. Atmospheric Environment, 20, pp. 71–85.

Kelly, N.A., Ferman, M.A., and Wolff, G.T. (1986): The Chemical and Meteorological Conditions Associated With High and Low Ozone Concentrations in Southeastern Michigan and Nearby Areas of Ontario. Journal of the Air Pollution Control Association, 36, pp. 150–158.

Krzanowski, W.J. (1988): Principles of Multivariate Analysis: A User's Perspective. Clarendon Press, Oxford.

Langstaff, J.E. and Pollack, A.K. (1985): Meteorological Characterisation of High Ozone Levels: A Pilot Study of St. Louis, Mo. Systems Applications Inc., California.

Ludwig, F. and Shelar, E. (1978): Effects of Weather Fronts on Ozone Transport. In Air Quality Meteorology and Atmospheric Ozone, ASTM STP 653, edited by A.L. Morris and R.C. Barras, American Society for Testing Materials, pp. 389–406.

Pollack, A.K. (1986): Application of a Simple Meteorological Index of Ambient Ozone Potential to Ten Cities. Paper presented in 79th Meeting of APCA, Minneapolis, June 22–27, 1986.

SAS Institute Inc. (1985): SAS User's Guide: Statistics. Cary, NC: Author.

Vukovich, F.M. and Fishman, J. (1986): The Climatology of Summertime O_3 and SO_2 (1977–1981). Atmospheric Environment, 20, pp. 2423–2433.

Yap, D., Ning, D.T., and Dong, W. (1988): An Assessment of Source Contributions to the Ozone Concentrations in Southern Ontario, 1979–1985. Atmospheric Environment, 22, pp. 1161–1168.

Zimmerman, K.A., Tropp, R., and Barta, R. (1987): The Relationship of Weather Patterns of Ozone in the Houston, Texas Area. Paper presented in 80th Annual Meeting of Air Pollution Control Association, June 1987, New York.

Received January 1989
Revised December 1989