# Statistical Models for Predicting Air Pollution in Taipei

GONG-YUH LIN AND LEE-CHIU HUANG

Reprinted from Proceedings of the National Science Council
Part A: Physical Science and Engineering
Vol. 9, No. 1, pp. 47-59, January 1985

National Science Council
Taipei, Taiwan
Republic of China

*Proc. Natl. Sci. Counc. ROC*(A)
Vol. 9, No. 1, 1985. pp. 47-59

# Statistical Models for Predicting Air Pollution in Taipei

GONG-YUH LIN* AND LEE-CHIU HUANG**

**Department of Geography*
*California State University*
*Northridge, California*
*U.S.A.*
***Department of Geography*
*National Taiwan University*
*Taipei, Taiwan*
*Republic of China*

(Received, 10 November 1983; Accepted, 26 October 1984)

ABSTRACT

The data of coefficient of haze observed in Taipei for the period 1973 through 1982 were analyzed to find diurnal, seasonal, and secular variations. The simple Markov chain probability model was used to explain the occurrences of polluted and unpolluted days. Multiple regression and discriminant models, using weather data and previous day's COH values as predictor variables, were developed. Maps of the correlation field showing the relationships between the strength of the northeast monsoon wind and the level of COH observed in Taipei were constructed.

*Key words: coefficient of haze (COH), simple Markov chain, multiple regression, discriminant model, polluted day, Chi-square test*

## Introduction

In a highly polluted urban area, a forecasting program is needed in order to implement pollution control rules and laws. For example, in Los Angeles an air pollution forecasting program has been functioning since 1955. The main purpose of pollution forecasting is to prevent excessive build-ups of atmospheric pollution so that the public health and property can be protected from the adverse effects of air pollution. Criteria of episodes (dangerous pollution levels) are predetermined and abatement actions, ranging from restricting all unnecessary driving to curtailing industrial and commercial operations, must be taken when the episode levels are either predicted or measured.

With a population of more than two million and rapid urbanization it is evident that Taipei is facing increasing air pollution emissions from automobile sources. Despite the efforts of the Bureau of Environmental Protection of Taipei (BEPT) to relocate the major steel manufacturers from the northeast sector of the city, i.e., Nankang and Sungshan, there are many small manufacturers in other areas of the city, including Sanchung and Panchiao. Under favorable weather conditions, pollution concentrations in Taipei exceed air quality standards, notably for particulate matters, NOx, and HC, at a few locations. Therefore, it is necessary to develop forecasting models for short-term (daily) enforcements of air pollution control laws.

In general, air pollution prediction models can be divided into two main categories: dispersion and statistical models. Dispersion models are derived from the principles of atmospheric circulations in the boundary layer. They are very costly and their accuracy needs improvement. On the other hand, statistical models are inexpensive and easy to develop and thus suitable for routine forecasting purposes. In the well-known smog city of Los Angeles, statistical models have been developed to forecast daily smog concentrations with a high degree of success.

## Methods

This research project has performed the following:

(1) The study of diurnal, seasonal, and annual variations of the coefficient of haze (COH) in Taipei. COH is a measure of reduced light transmission due to atmospheric particulates. It is determined by collecting particulates on a clean filter paper and then measuring the decrease of light transmission.

(2) The construction of synoptic weather maps of high and low COH levels for Taipei.

(3) The basic statistical analysis of COH data and the relationship to selected weather variables.

(4) The development of simple Markov chain probability models for COH levels at Taipei.

(5) The development of regression and discriminant models for predicting COH levels or categories using weather variables combined with the COH levels on the

previous day as predictor variables.

The SPSS [8] library program, available from the Treasury Department of the Executive Yuan, was employed to carry out statistical analyses. Daily COH data were derived from the Central Weather Bureau, the only location in Taipei where routine observations of COH levels are available over a long period of time. SC2 data available for this locations were used to develop regression models [1]. Data of other pollutants of interest are not available. Therefore, this research project has focused on the study of COH values only. Although the BEPT has published the monthly summary data of various kinds of pollutants observed at a few air monitoring stations, daily and hourly data are unavailable for performing statistical analyses in detail. Since July of 1982, the Bureau of Environmental Protection of the Department of Health (BEPDH) has established three air monitoring stations at Nankang, Sanchung and Panchiao to measure hourly concentrations of various types of pollutants. However, the data period at the time this research project started was too short to perform meaningful statistical analysis

Three sets of weather data were considered as predictor variables for developing regression and discriminant models with morning and afternoon peak mean hourly values of COH as predictors:

(1) The surface pressure, 1000- and 850-mb geopotential heights; the temperature, relative humidity, wind direction and speed at the surface, 1000- and 850-mb levels (18 variables in total).

(2) The lifting condensation level, level of free convection, convectional temperature, K-index, total index, maximum mixing height and temperature at the maximum mixing height derived from both Taoyuan and Taipei radiosonde data, and the maximum surface temperature at Taipei (12 variables in total).

(3) Surface pressure at 110 locations (110 variables) in Asia observed at $00_Z$ (8 a.m., Taipei local time).

The first two sets of weather data were derived from radiosonde observations over Taipei (Panchiao) at $1200_Z$ (8 p.m., Taipei local time) for the period June 1 through November 30, 1982. The maximum mixing height was calculated from radiosonde data observed at 8 a.m. over Taoyuan and at 8 p.m. over Taipei and the daily maximum surface temperature at Taipei. Since the autocorrelation coefficient of COH for lag one is approximately 0.40, the COH levels (mean, morning and afternoon peak hourly values) on the previous day were considered as part of the predictor variables for regression and discriminant models (Table 1).

Table 1. Variable List

| Codes | Variables |
|---|---|
| FCA: | Daily mean COH values on the first day (cohs/1000ft) |
| FC1: | Peak mean hourly COH values in the morning on the first day. |
| FC2: | Peak mean hourly COH values in the afternoon on the first day. |
| SC1: | Peak mean hourly COH values in the morning on the second day. |
| SC2: | Peak mean hourly COH values in the afternoon on the second day. |
| SRH: | Surface relative humidity (%). |
| U10: | 1000-mb relative humidity (%). |
| SDD: | Surface wind direction (degrees). |
| SFF: | Surface wind speed (m/sec). |
| STD: | 1000-mb dew point temperature (°C). |
| D10: | 1000-mb wind direction (degrees). |
| F10: | 1000-mb wind speed (m/sec). |
| T85: | 850-mb temperature (°C). |
| D85: | 850-mb wind direction (degrees). |
| F85: | 850-mb wind speed (m/sec). |
| LCL: | Lifting condensation level (mb). |
| LFC: | Level of free convection (mb). |
| MMT: | Maximum mixing height temperature derived from Taoyuan radiosonde data and maximum temperature at Taipei. |
| TMA: | Maximum daily temperature at Taipei (°C). |
| KIX: | K-index. |
| TIX: | Total index. |
| B: | Surface pressure at 53845. |
| C: | Surface pressure at 52889. |
| E: | Surface pressure at 47927. |
| I: | Surface pressure at 47772. |
| L: | Surface pressure at 57067. |
| M: | Surface pressure at 57036. |
| S: | Surface pressure at 57265. |
| U: | Surface pressure at 58238. |
| X: | Surface pressure at 58457. |
| Y: | Surface pressure at 58424. |

## Spatial and Temporal Variations

Table 2 illustrates 24-hour mean concentrations of different types of pollutants at various air monitoring stations in Taipei in 1982. It shows that the COH values reached or exceeded the unacceptable levels (2 cohs/1000 ft or higher) for the industrial areas including Nankang and Sungshan, and for the heavy traffic locations near the railroad station. Aside from the CO concentration in a subway on Linshen South Road, the concentrations of gaseous pollutants (CO, $SO_2$, $NO_2$ and $O_3$) were lower than the air quality standards for all air monitoring stations operated by BEPT. However, the pollution data measured at the three BEPDH stations reveals that NOx, HC, and $SO_2$ frequently exceeded the air quality standards for the period July through December 1982.

Mean hourly COH values show pronounced diurnal variations peaking at 8 a.m. and 11 p.m. in July and at 8 a.m. and 7 to 11 p.m. in January, with morning peak values slightly higher than the evening values (Figs. 1 and 2). Two peak values of COH on one day reflect the effects of heavy traffic hours and atmospheric stability. The evening rise of COH is associated with the increased thermal stability and lower wind speeds (Table 3).

It can be seen from Fig. 3 that COH values show

Table 2. Mean 24-hour Concentrations of Various Pollutants in Taipei, 1982

| | COH cohs/1000ft | CO ppm | $SO_2$ ppm | $NO_2$ ppm | HC ppm | $O_3$ ppm |
|---|---|---|---|---|---|---|
| Sungshan | 3.2* | 1.9 | 0.026 | | | |
| Nankang | 2.1* | 2.1 | 0.012 | | | 0.015 |
| Shihlin | 3.1* | 1.9 | 0.013 | | | |
| Chengchung | 3.3* | 2.8 | 0.031 | 0.019 | 3.68* | 0.008 |
| Chungshan | 2.8* | | 0.013 | 0.007 | 4.49* | |
| Peitou | 1.6 | 1.3 | 0.010 | | 2.77* | 0.011 |
| Suangyuan | 1.8 | 2.2 | 0.022 | 0.012 | 3.85* | 0.018 |
| Chingmei | 1.4 | 0.6 | 0.012 | 0.016 | | 0.025 |
| Neihu | 1.6 | 0.8 | 0.012 | 0.013 | | |
| Kuting | 0.1 | 1.6 | 0.013 | | | |
| Kwanchien | 6.1* | 5.5 | 0.044 | | | |
| Chungshan N. Road | 2.8* | 5.8 | 0.020 | | | |
| Chungching N. Road | 5.3* | 5.6 | 0.020 | 0.033 | | |
| Linshen S. Rd. subway | | 17.7* | | | | |
| Air Quality Standards | 2.0 24-hour average | 10.0 24-hour average | 0.10 24-hour average | 0.05 24-hour average | 0.50 24-hour average | 0.12 1-hour average |

* Pollution concentrations exceeded the air quality standards.

Table 3. Mean Hourly Wind Directions (Degrees) and Speeds (DM/S) Observed at the Three BEPDH Stations, October 1982

Wind Directions

| Station Hour: | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 Nankang | 158 | 147 | 175 | 156 | 147 | 156 | 152 | 156 | 171 | 182 | 168 | 169 | 188 | 177 | 175 | 140 | 137 | 148 | 169 | 141 | 152 | 170 | 152 | 145 |
| 2 Panchiao | 210 | 223 | 225 | 219 | 215 | 229 | 230 | 198 | 224 | 231 | 245 | 245 | 221 | 204 | 217 | 190 | 210 | 216 | 224 | 210 | 206 | 210 | 209 | 207 |
| 3 Sanchung | 249 | 247 | 247 | 223 | 234 | 230 | 207 | 221 | 231 | 252 | 231 | 233 | 216 | 215 | 211 | 222 | 227 | 238 | 237 | 240 | 251 | 249 | 249 | 257 |
| Ave: | 205 | 206 | 216 | 199 | 198 | 205 | 196 | 192 | 209 | 222 | 215 | 216 | 208 | 199 | 201 | 184 | 191 | 201 | 210 | 197 | 203 | 210 | 203 | 203 |

Wind Speeds

| Station Hour: | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 Nankang | 18 | 18 | 17 | 19 | 19 | 17 | 18 | 19 | 26 | 31 | 37 | 38 | 41 | 39 | 37 | 36 | 34 | 29 | 26 | 22 | 20 | 19 | 19 | 18 |
| 2 Panchiao | 9 | 9 | 9 | 9 | 9 | 8 | 9 | 10 | 11 | 12 | 13 | 15 | 16 | 16 | 17 | 16 | 14 | 12 | 11 | 10 | 9 | 9 | 9 | 10 |
| 3 Sanchung | 18 | 18 | 17 | 17 | 17 | 18 | 18 | 20 | 22 | 26 | 28 | 32 | 32 | 34 | 34 | 33 | 31 | 28 | 26 | 24 | 23 | 22 | 21 | 20 |
| Ave: | 15 | 15 | 14 | 15 | 15 | 14 | 15 | 16 | 19 | 23 | 26 | 28 | 30 | 30 | 29 | 29 | 26 | 23 | 21 | 19 | 17 | 17 | 17 | 16 |

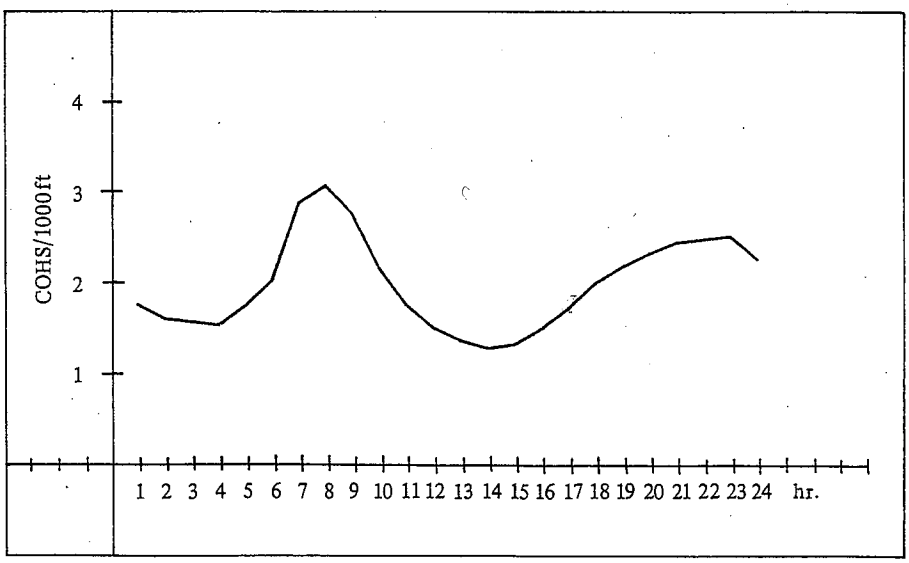


Fig. 1. Diurnal variations of COHS in July.

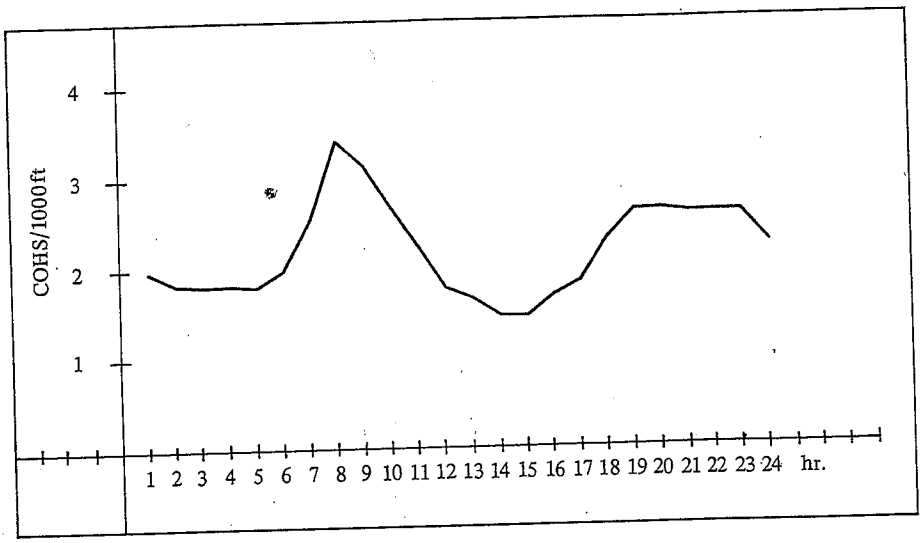


Fig. 2. Diurnal variations of COHs in January.

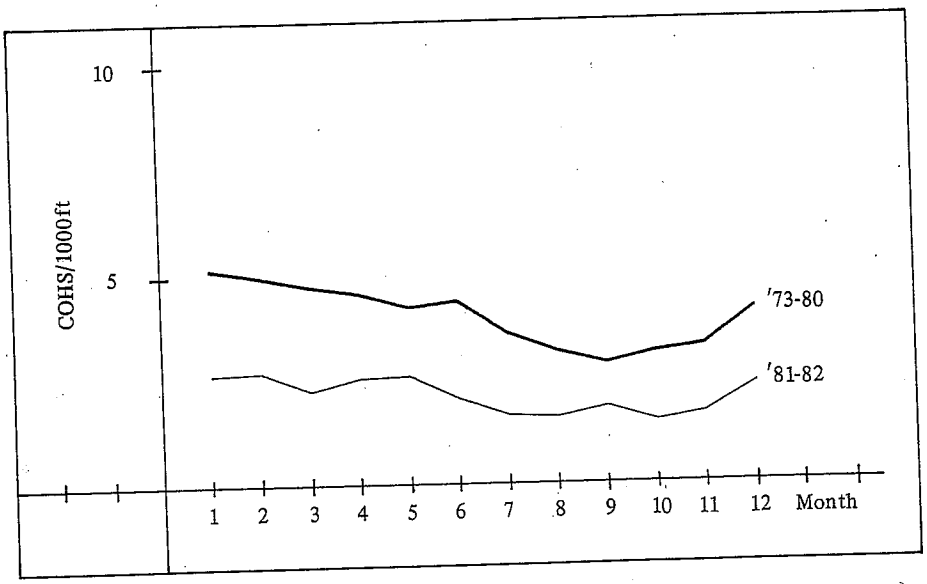


Fig. 3. Monthly variations in COHs for two different periods.

distinct seasonal variations with higher values in winter that is summer. In general, the ability of the atmosphere to dilute air pollutants is determined by two factors. The vertical dispersion capacity is hampered by the presence of an inversion layer. The horizontal dispersion capacity is mainly determined by wind speed. In Taipei, the inversion layer occurs more frequently and the inversion base is lower in winter than in summer (Table 4) because of more frequent occurrences of the Mongolian anticyclone in winter. Therefore, the atmosphere over Taipei is more stable in winter, limiting the vertical dispersion of pollution. In addition, Taipei is located in the downwind receptor area of the major emission source areas, Nankang and Sungshan, when the northeast wind prevails in winter.

There has been a significant decline in mean annual COH values since 1973; the annual mean decreases from

Table 4. Mean Monthly Values of Inversion Parameters Observed at 8 A.M. over Taipei, 1980

| Months | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Inversion Base Height (M) | 1610 | 1566 | 1326 | 1576 | 1504 | 1890 | 1811 | 2241 | 2353 | 1535 | 1444 | 1165 |
| Inversion Top Height (M) | 2263 | 2479 | 2119 | 2119 | 2176 | 2539 | 2867 | 2932 | 3273 | 2377 | 2104 | 1531 |
| Inversion Magnitude (°C) | 2.7 | 4.1 | 4.0 | 2.2 | 3.5 | 2.9 | 5.7 | 3.7 | 5.0 | 3.4 | 2.9 | 1.7 |
| Inversion Days | 23 | 14 | 28 | 20 | 14 | 3 | 7 | 7 | 8 | 12 | 12 | 15 |

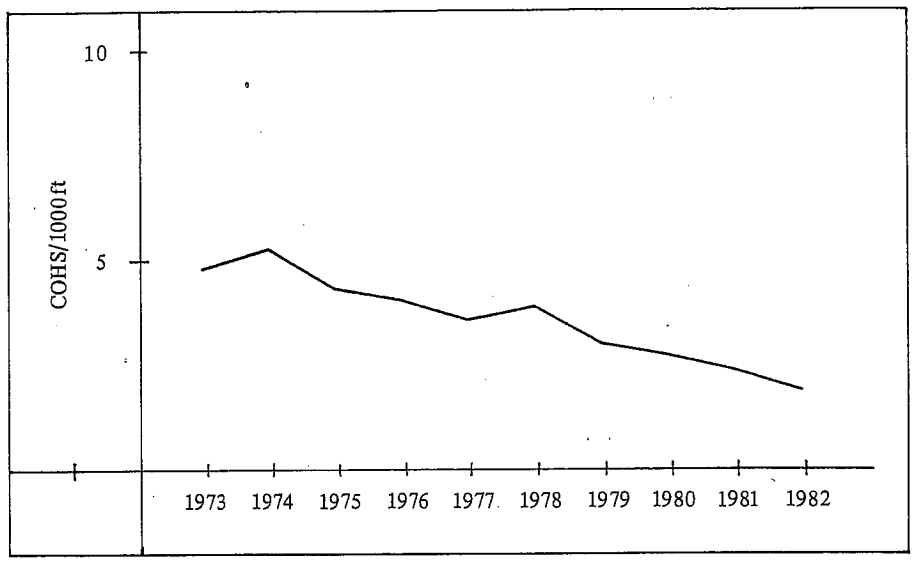


Fig. 4. Mean annual variations of COHS.

approximately 5 cohs/1000ft in 1973 to 2 cohs/1000ft in 1982, a 60% reduction (Fig. 4). Mean monthly and hourly values calculated for the period 1981-82 are considerably lower than those for the period 1973-80 (Figs. 3 to 6). The pronounced secular decline in COH levels is attributable to a series of legislative actions implemented since 1969 to reduce emissions of pollutants from both the stationary and automobile sources. Besides, an effort was made by the BEPT to relocate major steel manufacturers in Nankang and Sungshan to a newly developed industrial area, Tayuan (Taoyuan County), located to the south of Taipei. By 1981, seven of the eight major steel manufacturers have moved from Nankang and Sungshan, therefore, reducing the pollution levels in Taipei.

## Basic Statistics

For the study period June 1 through November 30, the mean, median, and mode of the peak hourly COH values in the morning are almost equal to each other; they are 1.5, 1.3, and 1.3, respectively (Table 5). These three values and a skewness of 0.85 as well as the shape of the histogram (Fig. 7) suggest that the frequency distribution of COH values is a positive skewness with more cases occurring in the values lower than the mode. The kurtosis of 1.87 illustrates that the frequency distribution is leptokurtic; the distribution has a peak values higher than the normal distribution and it has more extreme values. The cumulative frequency distribution curve indicates

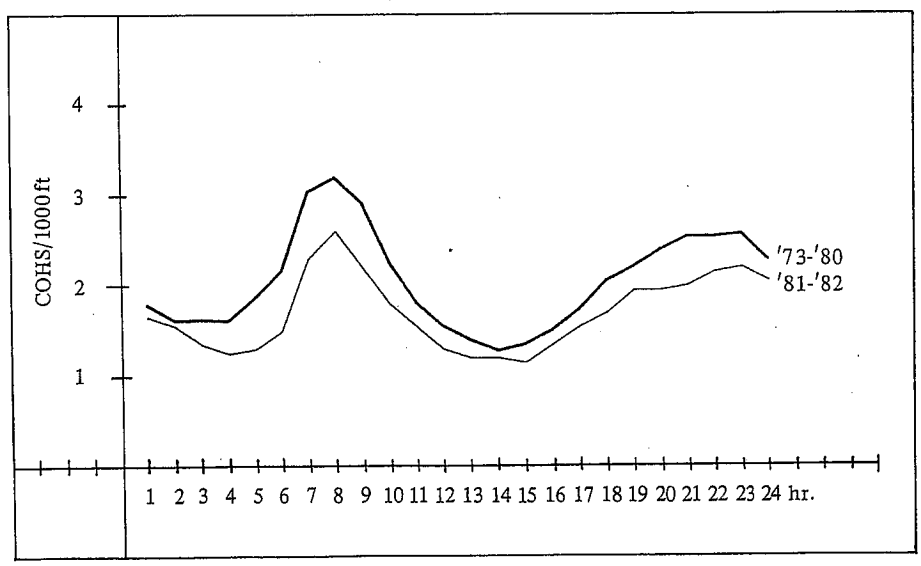


Fig. 5. Diurnal variations in COHS for two different periods in July.

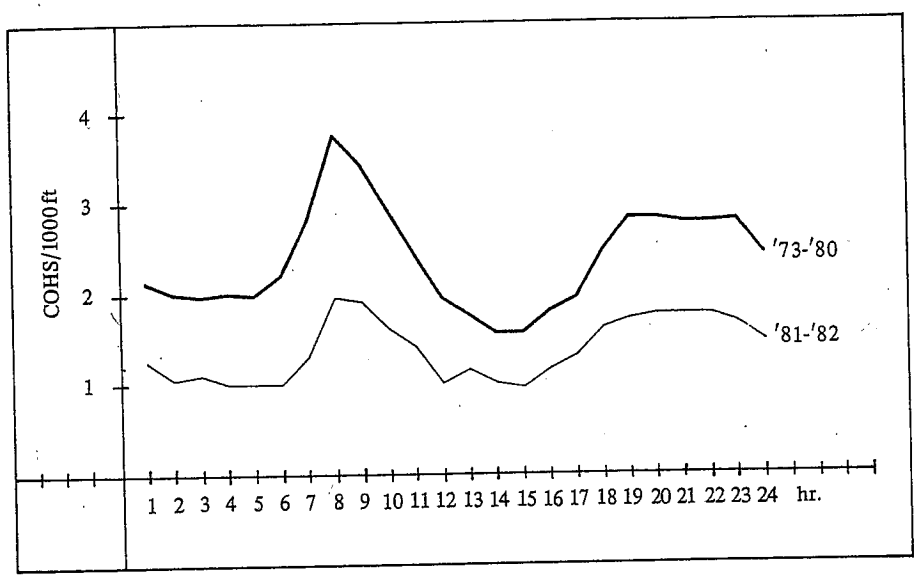


Fig. 6. Diurnal variations in COHS for two different periods in January.

that the COH values vary from 0.4 to 3.7 with a range of 3.3 (Fig. 8). The shapes of histogram and cumulative frequency distribution of peak mean hourly values in the afternoon are very similar to those of the morning COH values. The mean, median, and mode are 1.3, 1.3, and 1.1, respectively, reflecting the fact that the frequency distribution is also a positive skewness (0.76). The kurtosis of 0.83 proves that the frequency distribution is also leptokurtic. The COH levels vary from 0.4 to 3.3 with a range of 2.9. Therefore, statistical estimators of the morning COH levels are higher than those of the afternoon COH levels.

Table 5. Basic Statistical Characteristics for the Morning Peak Mean Hourly COH Values

| | | | | | |
|---|---|---|---|---|---|
| Mean | 1.455 | Std. Err. | 0.041 | Median | 1.336 |
| Mode | 1.300 | Std. Dev. | 0.539 | Variance | 0.291 |
| Kurtosis | 1.865 | Skewness | 0.852 | Range | 3.300 |
| Minimum | 0.400 | Maximum | 3.700 | | |

The autocorrelation coefficients for various lags indicate that the morning COH levels are more persistent than the afternoon COH levels. It appears that the COH

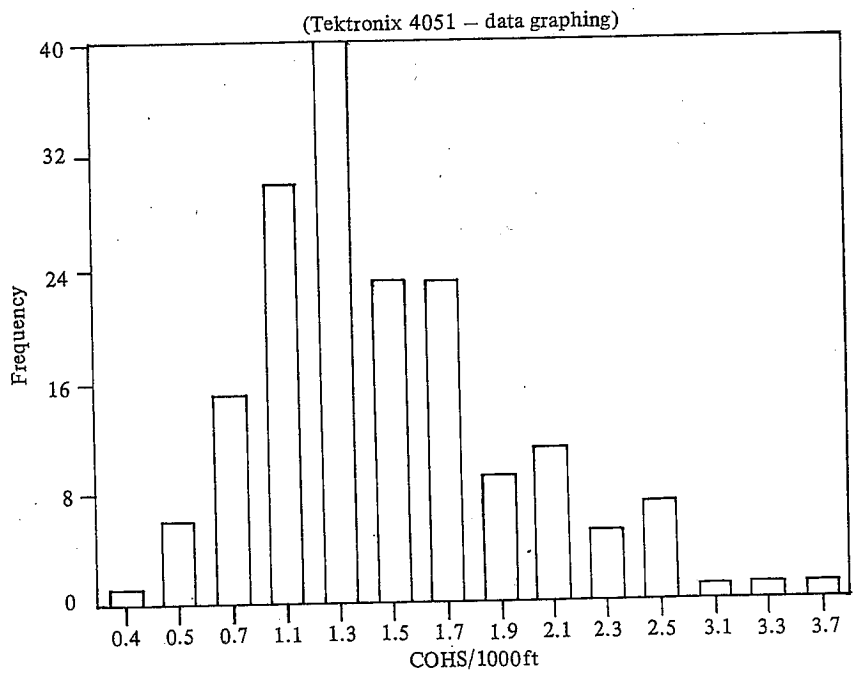


Fig. 7. Histogram of peak mean hourly COH values in the morning.

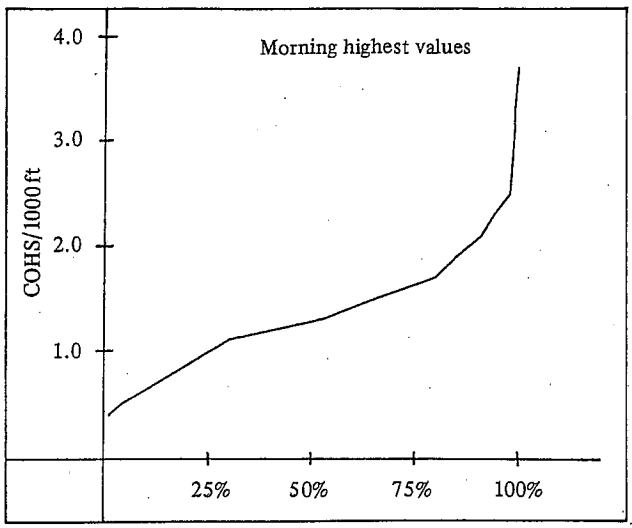


Fig. 8. Cumulative frequency distribution of peak mean-hourly COH values in the morning.

values, both in the morning and in the afternoon, fluctuate in a period of seven days as evidenced by the increase in the autocorrelation coefficients for the lag seven (Table 6).

Although a considerable number of weather variables taken from the radiosonde data over Taipei were correlated with the COH values on the second day, only a few of those variables show significantly high correlations (Table 7). Wind speeds at the surface and 1000-mb levels appear to have the highest correlation coefficients with the COH values. The correlation coefficients are −0.45 between the morning COH values and the 1000-mb wind speed and −0.32 between the afternoon COH values and the 1000-mb wind speed. These two correlation coefficients exceed the table value of −0.19 at the 1% significant level with 171 degrees of freedom (173 cases were used for calculating the correlations). Therefore, the correlation coefficients are significantly high to cause the rejection of the null hypothesis that there is no correlation between the COH values and wind speed at the 1000-mb level. In other words, there is a significant correlation coefficient between these two variables. Fig. 9 shows the correlation coefficients between the morning (8 a.m.) pressures at various weather stations in East Asia and the morning COH values observed on the second day in Taipei. The pressure at stations in Mongolia and northern China correlate most significantly high with the COH values in Taipei; the correlation coefficients slightly exceed −0.40. A negative correlation field north of Taiwan is observed while a positive field south of Taiwan is evident. This suggests that the weakening of the Mongolian

Table 6. Autocorrelation Coefficients for FC1 and FC2

Autocorrelation Function for Variable FC1

| Lag | Auto. Corr. | Stand. Err. | −1 −.75 −.5 −.25 0 .25 .5 .75 1 |
|---|---|---|---|
| 1 | 0.369 | 0.073 | . : . * |
| 2 | 0.213 | 0.073 | . : .* |
| 3 | 0.226 | 0.073 | . : . * |
| 4 | 0.234 | 0.072 | . : . * |
| 5 | 0.167 | 0.072 | . : * |
| 6 | 0.102 | 0.072 | . :*. |
| 7 | 0.192 | 0.072 | . : .* |
| 8 | 0.172 | 0.072 | . : * |
| 9 | 0.094 | 0.071 | . :*. |
| 10 | 0.078 | 0.071 | . :*. |
| 11 | 0.007 | 0.071 | . * . |
| 12 | −0.036 | 0.071 | .*: . |
| 13 | −0.004 | 0.071 | . * . |
| 14 | 0.026 | 0.070 | . :*. |
| 15 | 0.006 | 0.070 | . * . |
| 16 | −0.048 | 0.070 | .*: . |
| 17 | 0.021 | 0.070 | . * . |
| 18 | 0.070 | 0.070 | . :*. |
| 19 | 0.031 | 0.069 | . :*. |
| 20 | 0.110 | 0.069 | . :*. |

Autocorrelation Function for Variable FC2

| Lag | Auto. Corr. | Stand. Err. | −1 −.75 −.5 −.25 0 .25 .5 .75 1 |
|---|---|---|---|
| 1 | 0.354 | 0.073 | . : . * |
| 2 | 0.155 | 0.073 | . : * |
| 3 | 0.179 | 0.073 | . : .* |
| 4 | 0.066 | 0.072 | . :*. |
| 5 | 0.020 | 0.072 | . * . |
| 6 | 0.078 | 0.072 | . : *. |
| 7 | 0.091 | 0.072 | . : *. |
| 8 | 0.168 | 0.072 | . : * |
| 9 | 0.189 | 0.071 | . : .* |
| 10 | 0.018 | 0.071 | . * . |
| 11 | −0.018 | 0.071 | . * . |
| 12 | 0.014 | 0.071 | . * . |
| 13 | 0.074 | 0.071 | . :*. |
| 14 | 0.035 | 0.070 | . :*. |
| 15 | −0.055 | 0.070 | .*: . |
| 16 | 0.021 | 0.070 | . * . |
| 17 | 0.086 | 0.070 | . : *. |
| 18 | 0.137 | 0.070 | . : * |
| 19 | 0.091 | 0.069 | . : *. |
| 20 | 0.091 | 0.069 | . : *. |

* Autocorrelations
. Two Standard Error Limits

Table 7. Correlation Coefficients Between the COH Values and the Selected Weather Variables*

| Weather Variables | SRH | SDD | SFF | U10 | STD | D10 | F10 | T85 | D85 | F85 |
|---|---|---|---|---|---|---|---|---|---|---|
| SC1 | 0.03 | 0.20 | −0.44 | 0.02 | 0.08 | 0.26 | −0.45 | 0.21 | 0.31 | −0.33 |
| SC2 | 0.15 | 0.30 | −0.35 | 0.15 | 0.10 | 0.21 | −0.32 | 0.15 | 0.20 | −0.19 |

* Refer to Table 1 for the symbols of variables.

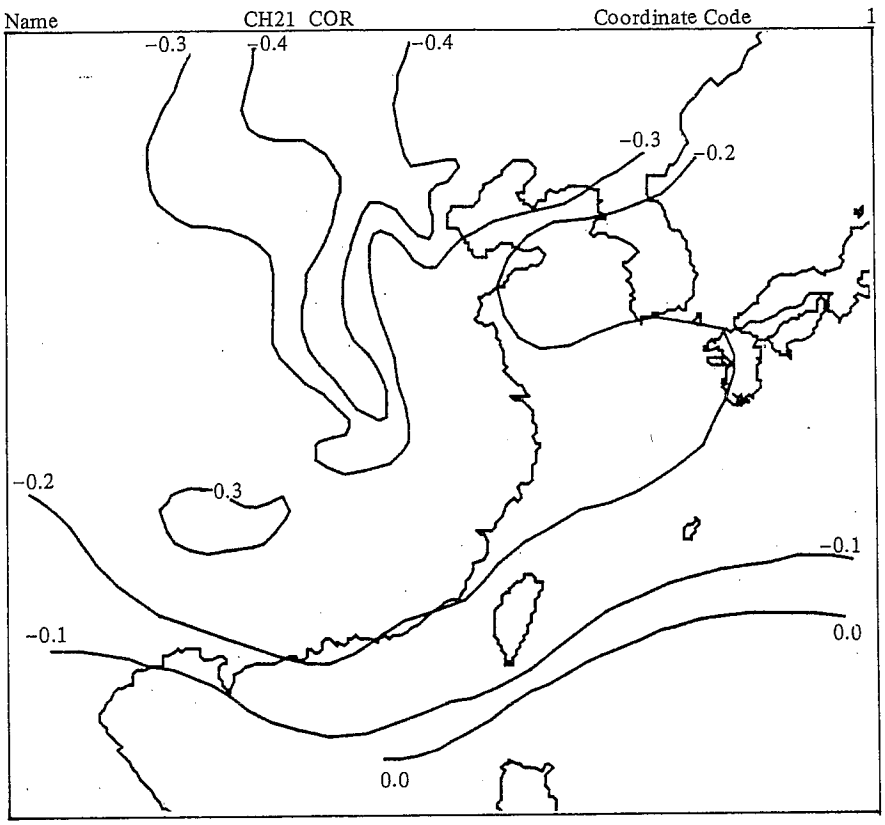


Fig. 9. Correlation field between the morning pressure and the morning highest COH value on the second day.

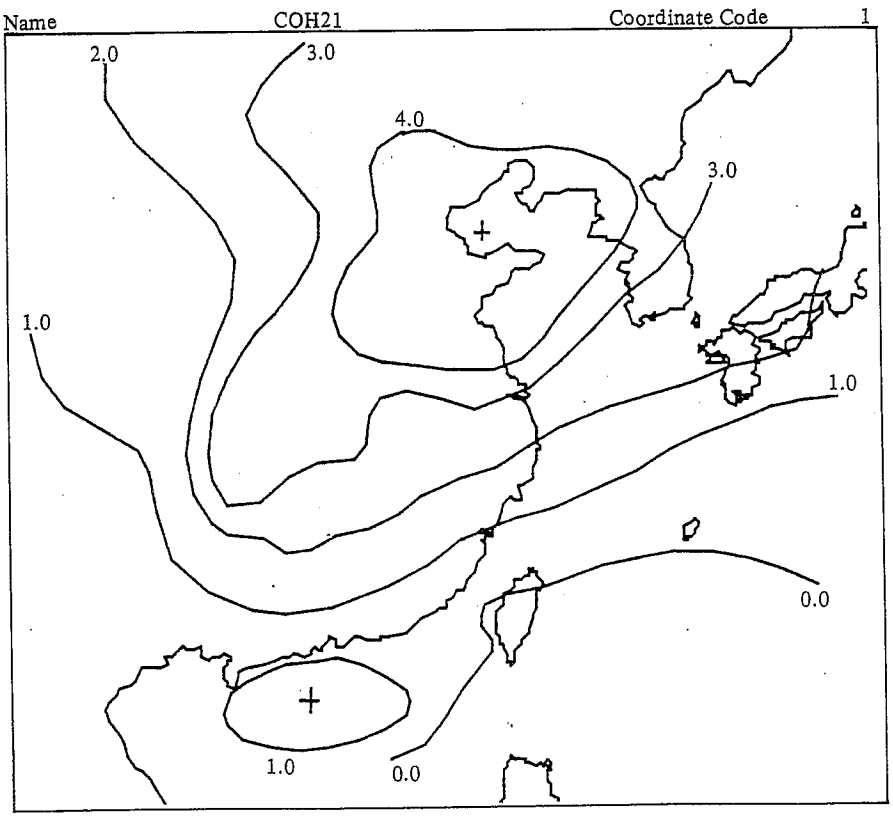


Fig. 10. Pressure differences between the low and high COH values in the morning on the second day.

anticyclone favors the build-up of COH values in Taipei as a result of the decrease in the strength of the northeast monsoon wind. The pressure patterns of the polluted morning (peak mean hourly COH values equaling or exceeding 2 cohs/1000ft) are not significantly different from those of the unpolluted morning. Generally, mean pressure are higher in the unpolluted morning by approximately 4 mb in northern China (Fig. 10).

## Simple Markov Chain Models

The first order Markov chain probability model has been employed frequently to study the occurrence of wet and dry spells in many locations [4, 10] and smog episodes in Los Angeles [6]. A finite Markov chain process is a finite stochastic process such that the outcome of any future trial is determined by the outcome of the last experiment and is independent of any previous outcome [5]. In this study, peak mean hourly COH values were divided into two states; the polluted and unpolluted states for both the morning and afternoon data, breaking at 2 cohs/1000ft. Two estimators, $P_1$ and $P_0$, must be calculated in order to perform the two-state Markov chain analysis. $P_0$ is the conditional probability of a polluted state given that the previous state was unpolluted (N), and $P_1$ is the conditional probability of a polluted state given that the previous state was also polluted (Y). According to the simple Markov chain probability model, any given morning or afternoon would fall into one of the following four categories, i.e., Y|Y, Y|N, N|Y, N|N. The first letter represents today and the second letter yesterday. $P_1$ and $P_0$ were derived from the following equations:

$$P_1 = \frac{n(Y|Y)}{n(Y|Y) + n(N|Y)},$$

$$P_0 = \frac{n(Y|N)}{n(Y|N) + n(N|N)},$$

where n denotes the total number of days for a given state. The probability of a polluted sequence of length k is

$$(1 - P_1)P_1^{k-1},$$

and the probability of a polluted sequence greater than k is

$$P_1^k.$$

Likewise, the probability of an unpolluted sequence of length k is

$$P_0(1 - P_0)^{k-1},$$

and the probability of an unpolluted sequence greater than k is

$$(1 - P_0)^k.$$

The mean first passage time or the average length of unpolluted spells is

$$P_0^{-1},$$

and the mean length of polluted spells is

$$(1 - P_1)^{-1}.$$

The estimated transitional probability matrix is

$$M = \text{(Yesterday)} \begin{matrix} \\ N \\ Y \end{matrix} \begin{matrix} \text{(Today)} \\ N \quad\quad Y \end{matrix} \begin{pmatrix} 1 - P_0 & P_0 \\ 1 - P_1 & P_1 \end{pmatrix}.$$

The equilibrium probability matrix derived from the k-step transitional probability matrix is

$$M^k = \begin{pmatrix} \pi_0 & \pi_1 \\ \pi_0 & \pi_1 \end{pmatrix}$$

where $\pi_0$ and $\pi_1$ are equilibrium or unconditional probabilities of unpolluted and polluted states, respectively.

Table 8 shows that $P_0$ and $P_1$ are 0.48 and 0.16 respectively, for the morning COH data, and 0.39 and 0.13, respectively, for the afternoon data. The transitional probability matrice are

$$M_1 = \begin{pmatrix} 0.84 & 0.16 \\ 0.52 & 0.48 \end{pmatrix}$$

for the morning COH data and

$$M_2 = \begin{pmatrix} 0.87 & 0.13 \\ 0.61 & 0.39 \end{pmatrix}$$

for the afternoon COH data. The 7th power of the transitional probability matrix for the morning COH data is

$$M_1^7 = \begin{pmatrix} 0.77 & 0.24 \\ 0.77 & 0.24 \end{pmatrix}$$

and the 6th power of the transitional probability matrix for the afternoon COH data is

$$M_2^6 = \begin{pmatrix} 0.82 & 0.18 \\ 0.82 & 0.18 \end{pmatrix}.$$

Table 8. Estimates of $P_1$ and $P_0$ for the Peak Mean Hourly COH Values Observed in the Morning (Upper) and the Afternoon (Lower)

| Yesterday | Today | | $P_1$ (Y\|Y) | $P_0$ (Y\|N) |
|---|---|---|---|---|
| | Yes | Yes & No | | |
| Yes | 38 | 79 | 0.48 | |
| No | 36 | 233 | | 0.16 |

| Yesterday | Today | | $P_1$ (Y\|Y) | $P_0$ (Y\|N) |
|---|---|---|---|---|
| | Yes | Yes & No | | |
| Yes | 22 | 57 | 0.39 | |
| No | 34 | 257 | | 0.13 |

Table 9. Predicted and Observed Frequencies of Polluted (Upper) and Unpolluted (Lower) Morning and Afternoon

| Lengths (days) | Morning | | Lengths (days) | Afternoon | |
|---|---|---|---|---|---|
| | PRI | OBS | | PRI | OBS |
| 1 | 20 | 19.7 | 1 | 18 | 20.3 |
| 2 | 10 | 9.5 | 2 | 15 | 12.7 |
| 3 | 8 | 8.8 | | | |
| chi-square | 0.10 | | | 0.65 | |
| Lengths (days) | Morning | | Lengths (days) | Afternoon | |
| | PRI | OBS | | PRI | OBS |
| 1 | 10 | 6.1 | 1 | 6 | 4.5 |
| 2 | 7 | 5.1 | 2 | 7 | 3.8 |
| 3 | 5 | 4.3 | 3-5 | 6 | 8.8 |
| 4-7 | 5 | 11.5 | 6-10 | 7 | 8.5 |
| 7 | 12 | 12.1 | 10 | 8 | 8.3 |
| chi-square | 7.08 | | | 4.30 | |

The occurrence of a polluted or unpolluted state on a given day is independent of the probability of the occurrence of any state seven days earlier for the morning COH and six days earlier for the afternoon COH.

The analysis of the mean first passage time indicates that it takes, in average, 6.5 days for the recurrence of the polluted morning but only 1.9 days for the recurrence of the unpolluted morning. The recurrence period of the polluted afternoon is 7.6 days while that of the unpolluted afternoon is 1.6 days. These reflect the fact that area near the station of the Central Weather Bureau is relatively unpolluted.

Chi-square tests were performed to approximate the observed frequences of sequences of polluted and unpolluted mornings or afternoons (Table 9). Calculated chi-square values are lower than table values at the 5% level, indicating that the Markov chain models approximate well the frequencies of the polluted and unpolluted spells.

## Regression Models

Stepwise multiple regressions were used to find the best linear prediction equations for peak mean hourly COH values in the morning and in the afternoon. Predictor variables were derived from the radiosonde data observed over Panchiao at 8 p.m. on the previous day, the first set of weather variables mentioned in the earlier section. Daily mean and the peak morning and afternoon COH values on the previous day were included as part of predictor variables. The regression equations thus provide approximately 8 hours in advance for predicting the morning COH values and 18 hours ahead for predicting the afternoon peak COH values. In the stepwise regression method, the predictor variable which has the highest correlation coefficient with the predictor enters the equation first. The second variable involved in the equation is selected from one of the remaining predictor variables that has the highest incremental contribution of F-value, the ratio of between-groups to within-groups sums of squares [3]. The process repeats until the desired incremental F-value is obtained. The regression equation involving the 5 most important predictor variables for predicting the morning COH values is

$$SC1 = 1.425 + 0.313(FC2) - 0.043(F10) + 0.143(FC1) - 0.007(U10) + 0.001(D85).$$

Symbols of all variables in the equation are explained in Table 1. The multiple correlation (R), defined as the correlation between the observed and predicted values, is 0.56 which yields a coefficient of determination ($R^2$) of 0.32, indicating that 32% of the variance in the morning COH values is explained by the combination of the five predictor variables in the equation. The regression equation for the afternoon COH values is

$$SC2 = 0.523 - 0.052(SFF) + 0.169(FC1) + 0.001(SDD) + 0.001(D85) + 0.005(SRH).$$

The $R^2$ for this equation is approximately 0.20. Evidently, the combination of weather variables and COH values on the first day provides little predictive power for predicting the COH values on the second day. Besides, the predictive power decreases as the prediction length increases from approximately 8 hours to 18 hours.

Using the second set of weather variables, representing stability indice, mentioned in the earlier section, and the COH values as predictor variables, the regression equations are

$$SC1 = 0.789 + 1.177(FCA) + 0.245(FC2) - 0.002(LFC) - 0.153(FC1) - 0.010(MMT),$$

with $R^2$ equaling 0.41, and

$$SC2 = 1.041 + 0.282(FC2) + 0.459(FCA) - 0.003(LFC) + 0.013(KIX) - 0.016(TIX),$$

with $R^2$ equaling 0.26. Obviously, there are slight increases in predictive powers by using the second set of weather variables as part of the predictor variables for predicting SC1 and SC2.

Combining the COH levels and the third set of weather variables (surface pressures) as predictor variables, the regression equations are

$$SC1 = -2.061 + 0.460(FC2) - 0.003(L) + 0.129(X) - 0.123(U) + 0.589(FCA),$$

and

$$SC2 = 11.328 - 0.078(M) + 0.032(Y) + 0.018(C) + 0.156(FC1) + 0.018(B).$$

$R^2$ is 0.57 for the SC1 equation and 0.21 for the SC2 equation. There is a 16% increase in $R^2$ for predicting

SC1 from surface pressures over that from stability indice. However, there is a slight decrease of 5% in $R^2$ for predicting SC2 from surface pressures in place of stability indice.

## Discriminant Models

Discriminant analysis is a useful statistical tool for categorical forecasts. For instance, this technique was used to develop models for predicting oxidant episodes [7] and infrared visibility categories [3] from weather data. Mathematical treatments of discriminant analysis were discussed in detail by a number of multivariate statistical textbooks [2, 9]. The aim of discriminant analysis is to classify individual cases from their scores on discriminant functions which are linear combinations of a set of discriminating or predictor variables. A discriminant function is a vector which maximizes the ratio of between-groups to within-groups sums of squares of discriminant scores or F-ratio of variance estimate. The function is derived in a way to achieve not only the maximum group differentiations but also the minimum probability of errors in assigning cases to groups.

Two-group discriminant functions were sought for predicting COH values breaking at 2 cohs/1000ft. The standardized discriminant functions using the first set of weather variables as predictors are

$$SC1 = -0.522(STD) + 0.360(D85) + 0.431(FC1) + 0.626(FC2),$$

and

$$SC2 = 0.677(SRH) + 0.415(SDD) - 0.525(SFF) + 0.331(F85).$$

Data of discriminating variables should be normalized, i.e., z-scores with mean 0 and variance 1, in order to obtain the discriminant scores for individual cases from the standardized discriminant functions. The discriminant coefficients indicate the relative importance of each variable to the function. It can be seen that FC2 contributes the highest portion of discriminant scores for SC1 and SRH does the same for SC2. A discriminant score of a case is the number of standard deviations from the grand mean (zero discriminant score) of all observations on the discriminant function.

It is impractical to use standardized discriminant functions for classification or prediction purposes because values of predictor variables must be normalized first. Therefore, unstandardized discriminant functions are more convenient for prediction purposes:

$$SC1 = -0.295 - 0.159(STD) + 0.004(D85) + 0.845(FC1) + 1.476(FC2),$$

and

$$SC2 = -5.703 + 0.073(SRH) + 0.005(SDD) - 0.325(SFF) + 0.070(F85).$$

It is notable that a constant term is added into the unstandardized discriminant functions. Observations of predictor variables are applied to the equations without any transformations. For SC1, the centroid (group mean) is -0.190 for the group of COH < 2 and 1.364 for COH ≥ 2. The midpoint of group centroids is 0.587. If the calculated SC1 is equal to or greater than 0.587, the case is classified as having COH > 2, and vice versa. For SC2, the group centroids are -0.098 for COH < 2 and 1.398 for COH ≥ 2.

Prediction or classification can be achieved alternatively by using classification functions. For the morning COH levels, the classification functions are

$$SC1(0) = -22.266 + 1.814(STD) + 0.005(D85) + 2.751(FC1) + 0.528(FC2),$$

and

$$SC1(1) = -23.636 + 1.568(STD) + 0.012(D85) + 4.064(FC1) + 2.822(FC2).$$

If the calculated value of SC1(1) is greater than that of SC1(0), the case is classified as having COH ≥ 2, and vice versa. The overall accuracy of the classification using either the discriminant functions or classification functions is 77.3%. The classification functions for the afternoon COH data are

$$SC2(0) = -47.342 + 1.004(SRH) + 0.048(SDD) + 3.186(SFF) + 0.035(F85),$$

and

$$SC2(1) = -56.844 + 1.112(SRH) + 0.055(SDD) + 2.699(SFF) + 0.140(F85).$$

The overall accuracy of the classification is 74.8%. There is a slight decrease in the classification (prediction) accuracy as the prediction time extends from morning to afternoon. This is also reflected by a decrease of canonical correlation coefficients from 0.46 for SC1 to 0.35 for SC2, and an increase of Wilks' Lambda values from 0.79 for SC1 to 0.88 for SC2. Canonical correlation is the maximum correlation between the linear function of predictor variables and that of predictors that are coded 1 for COH ≥ 2 cohs/1000ft and 0 for COH < 2 cohs/1000ft. Wilks' Lambda, the ratio of the within-groups to the total sums of squares, is an inverse measure of discriminant power.

For simplicity, only unstandardized discriminant functions derived from other sets of weather variables are presented in the following. Based on the stability indice, the functions are

$$SC1 = -1.169 - 0.001(LFC) - 0.058(TMA) + 2.979(FCA) + 0.941(FC2) + 0.941(FC2),$$

with group centroids of -0.285 for COH < 2 and 1.515 for COH ≥ 2, and

$$SC2 = -23.422 + 0.021(LCL) - 0.002(FC) + 0.138(MMT) + 0.925(FC2) + 0.925(FC2),$$

with group centroids of -0.106 for COH < 2 and 1.174 for COH ≥ 2. The overall accuracy of predictions are 78.5% for SC1 and 75.5% for SC2. Evidently, there is a slight improvement in prediction power of these models over the preceding models. This is also reflected by higher canonical correlation coefficients of 0.55 for SC1 and 0.33 for SC2 derived from the stability indice.

Unstandardized discriminant functions using the third set of weather data (surface pressures) as discriminating variables are

$$SC1 = 2.659 - 0.55(I) + 0.049(S) + 0.948(FCA) + 1.850(FC2),$$

with group centroids of -0.310 for COH < 2 and 1.673 for COH ≥ 2, and

$$SC2 = -194.509 + 0.191(E) + 1.528(FC2),$$

with group centroids of -0.112 for COH < 2 and 1.406 for COH ≥ 2. The overall accuracy of classifications are 86.4% for SC1 and 80.2% for SC2. The canonical correlation coefficients are 0.59 for SC1 and 0.37 for SC2. The models derived from surface pressures show significant improvements over the preceding models derived from either the stability indice or radiosonde data.

## Summary

This study reveals that the mean hourly COH values observed at the station of the Central Weather Bureau show pronounced diurnal variations peaking at 8 a.m. and 7 through 11 p.m., conforming well with peak traffic hours and increasing stability of air in the evening. There has been a significant decline in the COH trend since 1973 following a series of legislative actions aiming at reducing pollution emissions, especially from the major stationary sources located in the northeast sector of the city, including Nankang and Sungshan.

Statistical analysis of COH data indicates that the COH level on the previous day is very important in determining the COH level today, as evidenced by an autocorrelation coefficient of approximately 0.40 for lag one either for the morning or afternoon data. Mean synoptic weather maps suggest that the mean pressure patterns for polluted and unpolluted periods are more or less the same; the Mongolian anticyclones were well defined for both periods. The only difference is that there is a decrease in the pressure gradient between North China and Taiwan for the polluted period. This reflects the fact that the weakening of the northeast wind promotes the COH levels in Taipei, as is also substantiated by a negative correlation coefficient of approximately -0.40 between the surface or 1000-mb wind speeds and the morning peak COH values.

The occurrences of polluted mornings and afternoons were found to follow the simple Markov chain process. It requires approximately 7 days for the recurrence of polluted mornings or afternoons, but only 2 days for the recurrence of unpolluted periods.

Three sets of weather variables, radiosonde observations at 8 p.m. over Taipei, stability indice derived from radiosonde data, and surface pressures of 110 weather stations in East Asia observed at 8 a.m., together with the COH levels on the first day were employed as predictor variables to develop regression and discriminant models for predicting the COH levels on the second day. It was found from the regression analysis that the weather variables have low predictive powers for COH levels. Values of $R^2$ vary from approximately 60% to 32% for different regression models predicting the morning COH levels and from 26% to 20% for predicting the afternoon COH levels. It can be concluded that regression models are of little use in predicting the COH levels. However, if categorical forecasts of COH levels breaking at 2 cohs/1000 ft are of interest, the discriminant models using three sets of weather variables, each combined with the COH levels on the previous day, provide a moderately high degree of accuracy for predictions. The accuracy of predictions or classifications varies from approximately 86% for predicting the morning COH categories to 75% for predicting the afternoon COH categories from the surface pressures at a number of weather stations in East Asia combined with the COH levels on the previous day. Obviously, the accuracy of predictions from both regression and discriminant models decreases as the prediction time extends from morning to afternoon.

## Acknowledgements

The author is Professor of Geography at California State University, Northridge. The research project was conducted when he was employed as visiting professor at National Taiwan University during the 1982-83 academic year. This work was undertaken under contract NSC 72-0202-M002-5 with the National Science Council of the Republic of China.

## References

1. Chang, Che-Ming. 1981. The assessment of sulfur dioxide pollution potential in Taipei. *Meteorological Bulletin* (Taiwan), Central Weather Bureau, 28(1), 41-62.
2. Cooley, W.E. and P.R. Lohnes. 1971. *Multivariate Data An-*

*alysis.* Wiley, New York.
3. Court, A., Lin, G-Y. and S. Zetsche. 1982. Infrared visibility prediction by statistical methods. *PAGEOPH* (Birkhauser Verlag, Basel), **120**, 203-210.
4. Gabriel, K.R. and J. Neumann. 1962. A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quarterly Journal* of the Royal Meteorological Society, 88, 90-95.
5. Kemeny, J.G. and L.L. Snell. 1960. *Finite Markov Chains.* New York: Van Nostrand.
6. Lin, G-Y. 1981. Simple Markov chain model of smog probability in the south coast air basin of California. *Professional Geographer*, AAG, 33, 228-236.
7. Lin, G-Y. 1982. Oxidant prediction by discriminant analysis in the south coast air basin of California. *Atmospheric Environment,* 16(1), 135-143.
8. Nie, N.H. *et al.* 1975. *Statistical Package for the Social Sciences* (*SPSS*), McGraw-Hill, New York.
9. Tasuoka, M.M. 1971. *Multivariate Analysis,* Wiley, New York.
10. Weiss, L.L. 1964. Sequences of wet or dry days described by a Markov chain probability model. *Monthly Weather Review,* 92, 169-176.

# 臺北市空氣污染預報之統計模式

林 功 豫* • 黃 麗 珠**

*美國加州大學地理系
**國立臺灣大學地理系

## 摘 要

本研究以中央氣象局測站觀測所得之煤塵濃度資料作統計分析，研究結果顯示臺北市煤塵濃度一日有二高峯，發生於上午 8 時及下午 7 至11時，此與交通量尖峯時間及入晚後空氣穩定度之增加有關。自民國62年以來，政府訂立並執行一連串之空氣污染防制法規，勸導 7 家主要鋼鐵工廠遷出南港和松山區，臺北市煤塵濃度有顯著減少之趨勢。

若以煤塵濃度達到或超過2 COHS/1000ft 爲污染期，則污染期或非污染期發生之頻率，不管是在上午或下午，皆可以簡易馬可夫連鎖模式解釋其發生之或然率。另以三組不同之觀測天氣因子配合第一天煤塵濃度之平均值，上午和下午之最高值作逐步廻歸和判別模式，以預報第二天上下午煤塵之尖峯濃度，結果顯示廻歸模式不適宜於預報之用。若以不同之判別模式對高煤塵濃度之發生與否作預報分析，則準確度可達 75至86%，其中以東亞某些測站上午 8 時之氣壓作爲部份預報變數，所得之預報結果最佳。平均綜觀東亞天氣圖顯示，在高低污染期，天氣系統無顯著變化，惟華北和臺灣間氣壓梯度減弱時，臺北煤塵濃度會增高，此與東北季風之減弱，因此減低了水平擴散速率有關。