MASTER

COO-2874-24

APPLICATION OF SOME ROBUST STATISTICAL PROCEDURES
TO AIR POLLUTION DATA IN THE SAN FRANCISCO BAY AREA.

by

CHEN-HUI FOO

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

No.
WORKING PAPER #6

This draft is intended for limited circulation.

September 1977

EB
DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

This research was supported by Grant from SIMS

Prepared For
THE U.S. ENERGY RESEARCH AND DEVELOPMENT ADMINISTRATION
UNDER CONTRACT NO. EY-76-S-02-2874

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Application of Some Robust Statistical Procedures
to Air Pollution Data in the San Francisco Bay Area

Chen-Hui Foo

1. Introduction.

In a paper by Cleveland and Guarino [1], the authors investigated production and origin of ozone within the New York City-northern New Jersey region. Biweight regression analysis was used to relate daily maximum ozone concentrations at pairs of sites, and the 10% trimmed means was used to examine the diurnal variability of concentration. From their results, it seems worthwhile to apply these methods to data from the Bay Area. However, we will use biweight regression analysis for a different purpose--to examine relationships between daily maximum hourly averages of total oxidant concentration and selected weather variables. The study of daily variations of NO_1, NO_2 and total oxidant by using the 10% trimmed mean is also included. The data base consists of the hourly concentrations of the pollutants at San Jose and Redwood City monitoring stations during July, August, and September in 1974. [2]

2. The relationship between daily maximum high hour oxidant concentration and the selected weather variables.

Figures 1, 2, and 3 show the scatter-plot relationship between daily maximum hourly average of oxidant concentration at San Jose and the mean wind speed, maximum temperature and stability factor, respectively. From the figures, it appears that daily maximum hourly average of oxidant concentration is positively correlated with daily maximum temperature and

stability factor. But the relationship between the oxidant concentration and the mean wind speed is not clear.

In order to further quantify the relationship, we regress the daily maximum oxidant concentration on the above three weather variables. Bi-weight regression fitting [3] is used here which is a special case of the weighted least squares method. Therefore, instead of doing usual regression analysis by putting the same weights on all days, we allow different days to have different variation by using biweight regression fitting. To compare these two methods, a detailed description of the analysis of the oxidant concentration at San Jose monitoring station will be given.

Figure 1

Figure 2

Figure 3

Let $\tilde{Y}$ be the column vector of daily maximum hourly average of oxidant concentration. X_1, X_2, X_3 are the mean wind speed, maximum temperature, and stability factor, respectively. The linear model is $\tilde{Y} = \beta_0 + \beta_1 \tilde{X}_1 + \beta_2 \tilde{X}_2 + \beta_3 \tilde{X}_3 + \tilde{\varepsilon}$ where $E(\tilde{\varepsilon}) = \tilde{0}$, $\text{Var}(\tilde{\varepsilon}) = \tilde{V} \delta^2$ and $\tilde{V}$ is positive definite. The usual least square method considers $\tilde{V}$ as $\tilde{I}$ and the fitted regression line is

$$\text{LOG}(\tilde{Y}) = -1.88 + 0.107 \tilde{X}_1 + 0.0431 \tilde{X}_2 + 0.0164 \tilde{X}_3$$

The plot of standardized residuals vs. predicted $\tilde{Y}$ is shown in Figure 4. The apparent outliers indicate that the maxima associated with those outliers have an undue influence on the determination of the fitted line.

In biweight regression fitting, $\tilde{V} = \tilde{W}^{-1}$ and

$$\tilde{W} = \begin{pmatrix} W_{11} & 0 & \dots & 0 \\ 0 & W_{22} & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & W_m \end{pmatrix} \text{ is the weight matrix.}$$

m = the total number of days observed.

Figure 4

Since V is unknown and depends on the data, usually it is necessary to make the assumption $V = I$ and then attempt to adjust V by examining the residuals from regression analysis. The iteration procedures are as follows,

$$\hat{\beta}_{n+1} = (\tilde{X}'_n W_n \tilde{X}_n)^{-1} \tilde{X}'_n W_n \tilde{Y}_n \quad n = 0, 1, 2, \dots$$

where

$$\begin{matrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{matrix} \quad \sim x = \begin{matrix} 1 & X_{11} & X_{21} & X_{31} \\ 1 & \cdot & \cdot & \cdot \\ \vdots & \cdot & \cdot & \cdot \\ 1 & X_{1m} & X_{2m} & X_{3m} \end{matrix}$$

X_{ik} = the observation of the i -th weather variable on
 k -th day

$$\tilde{W}_n = \begin{pmatrix} W_{11,n} & 0 & & \dots & 0 \\ 0 & W_{22,n} & 0 & \dots & 0 \\ \vdots & \vdots & & & \\ 0 & 0 & & \dots & W_{mm,n} \end{pmatrix} \quad \tilde{W}_0 = \tilde{I}$$

$$W_{jj,n} = w \left(\frac{y_j - (X_j \tilde{\beta}_n)}{c S_n} \right) \quad \text{for } n \geq 1$$

S_n = the inter-quartile range of residuals in n -th
iteration (I.Q. range in n -th step)

c is an arbitrary constant.

$$w(u) = \begin{cases} (1-u^2)^2 & \text{for } |u| \leq 1 \\ 0 & |u| > 1 \end{cases}$$

For a Gaussian random variable, the probability of being inside the
interval [lower quartile - 1.5 I.Q. range, upper quartile + 1.5 I.Q.
range] is 99 %.

Hence if we take $c = 2$, the probability that the residuals will
fall in [- 2 I.Q. range, 2 I.Q. range] is 99%.

The equivalent method used by Cleveland and Guarino [1] is similar to the above procedure except in the iteration procedure

$$\mu_{kk,n} = \begin{cases} \left[1 - \left(\frac{r_{k,n}}{6r'} \right)^2 \right]^2, & r_k < 6r' \\ 0 & , r_k > 6r' \end{cases}$$

$$w_{kk,n} = \frac{\mu_{kk,n}}{\sum_{j=1}^m \mu_{jj,n}} \quad (\text{normalized})$$

where

$r_{k,n}$ = the absolute value of the residual on k-th day in the n-th iteration step

r' = median of $\{r_k\}$

The fitted line from BIWEIGHT (biweight regression fitting) after the fifth iteration is

$$\text{LOG}(Y) = -1.82 + 0.077 \underline{x}_1 + 0.044 \underline{x}_2 + 0.015 \underline{x}_3$$

The explained percentage of the variation (R^2) increases from 49.5% to 66.8% . Figure 5 shows the standardized residual plot from BIWEIGHT. The standardized residuals are more evenly scattered around 0

and without strange outliers. During the iteration procedure, we consistently threw away one outlier on July 14, 1974 with an unusually low daily maximum hourly average of oxidant concentration - 1 pphm.

Figure 5

Advantages of using biweight function:

- (1) Placing lower weights on large residuals allows these larger residuals to be still larger.
- (2) Relaxing the constraints imposed by these large residuals ought to make the small residuals still smaller.

Therefore, if we put zero weights on some observations, the adjusted weights in the following iteration steps are always zero. Hence, we consistently threw away the outliers in each iteration procedure.

The comparison between the predictions from BIWEIGHT regression and the usual least square method is shown in Table 1. $\log Y_B$ indicates the former prediction and $\log Y_0$ indicates the latter prediction.

In the biweight regression analysis, we allow each individual observation to have different variance. But we do not exploit the statistical nonindependence of the observations. The short time lag effect of air pollutant concentration is generally quite evident.

TABLE 1

 X_1 = Mean Wind Speed (mph) X_2 = Maximum Temperature ($^{\circ}$ F) X_3 = Stability Factor

Date	X_1	X_2	X_3	$\log Y$	$\log \hat{Y}_0$	$\log \hat{Y}_B$
7/1	4.2	88	7	1.95	2.47	2.49
7/2	4.6	88	-1	2.71	2.38	2.40
7/3	4.4	85	7	2.77	2.37	2.37
7/4	2.5	89	11	3.09	2.4	2.46
7/5	2.5	85	7	1.79	2.16	2.22
7/6	3.3	76	-13	1.61	1.53	1.58
7/7	2.6	74	-11	1.61	1.40	1.48
7/8	4.5	71	-12	0.69	1.46	1.48
7/9	3.0	72	-11	0.69	1.36	1.42
7/10	2.7	75	-14	1.39	1.41	1.49
7/11	2.9	78	-6	1.61	1.69	1.75
7/12	2.8	84	2	1.79	2.07	2.13
7/13	2.5	82	5	1.95	2.00	2.06
7/14	2.4	89	10	0.0	2.37	2.44
7/15	3.0	79	4	1.39	1.91	1.95
7/16	3.4	78	-4	1.61	1.78	1.82
7/17	3.4	80	-1	1.39	1.91	1.95
7/18	3.2	88	4	1.61	2.32	2.36
7/19	3.2	91	12	2.48	2.58	2.62
7/20	3.5	87	14	2.56	2.47	2.50
7/21	3.2	88	13	2.20	2.46	2.50
7/22	3.2	91	8	2.30	2.51	2.56
7/23	3.0	92	15	2.83	2.65	2.69
7/24	3.7	100	15	3.18	3.07	3.10
7/25	3.2	97	5	2.71	2.72	2.78
7/26	4.5	93	5	2.71	2.69	2.70
7/27	4.1	81	0	2.30	2.05	2.07
7/28	3.6	79	6	1.95	2.00	2.03
7/29	3.7	80	11	1.95	2.14	2.16
7/30	3.5	80	17	1.95	2.22	2.23
7/31	3.2	81	5	1.95	2.03	2.07

TABLE 1 (cont.)

 X_1 = Mean Wind Speed (mph) X_2 = Maximum Temperature ($^{\circ}$ F) X_3 = Stability Factor

Date	X_1	X_2	X_3	$\log Y$	$\log \hat{Y}_0$	$\log \hat{Y}_B$
8/1	2.9	90	14	2.56	2.54	2.58
8/2	4.0	90	20	2.94	2.75	2.76
8/3	3.3	85	14	2.64	2.36	2.39
8/4	4.0	86	12	2.40	2.45	2.46
8/5	3.3	88	10	2.20	2.43	2.46
8/6	3.4	94	8	2.71	2.66	2.70
8/7	3.4	80	12	1.79	2.12	2.15
8/8	5.1	89	10	2.77	2.66	2.65
8/9	5.6	85	10	2.64	2.54	2.51
8/10	4.5	82	-2	2.64	2.10	2.11
8/11	3.8	84	9	2.56	2.29	2.31
8/12	5.5	84	-10	2.20	2.16	2.16
8/13	5.2	82	-3	1.79	2.16	2.15
8/14	4.0	88	10	2.71	2.50	2.52
8/15	4.6	78	6	2.20	2.07	2.06
8/16	3.3	76	-4	1.61	1.68	1.72
8/17	3.1	78	8	2.08	1.94	1.98
8/18	3.2	80	8	1.79	2.04	2.07
8/19	3.1	84	1	2.19	2.09	2.14
8/20	2.9	94	8	2.40	2.61	2.67
8/21	2.8	96	19	2.77	2.86	2.91
8/22	3.6	95	21	3.00	2.94	2.96
8/23	2.8	93	5	2.08	2.51	2.57
8/24	4.1	90	19	2.89	2.75	2.75
8/25	4.8	88	15	2.56	2.67	2.65
8/26	3.5	84	8	2.40	2.24	2.27
8/27	3.1	78	6	1.79	1.91	1.95
8/28	3.3	80	-2	1.61	1.88	1.93
8/29	3.2	78	-7	1.95	1.71	1.76
8/30	2.9	73	0	1.39	1.57	1.62
8/31	3.0	76	-3	1.79	1.66	1.72

TABLE 1 (cont.)

 X_1 = Mean Wind Speed (mph) X_2 = Maximum Temperature ($^{\circ}$ F) X_3 = Stability Factor

Date	X_1	X_2	X_3	log Y	$\log \hat{Y}_0$	$\log \hat{Y}_B$
9/1	3.0	79	8	2.08	1.97	2.01
9/2	3.0	80	11	2.20	2.07	2.10
9/3	2.9	80	5	2.20	1.96	2.00
9/4	2.4	88	8	2.48	2.30	2.36
9/5	2.3	98	16	3.14	2.85	2.92
9/6	2.7	95	10	3.04	2.66	2.72
9/7	3.1	86	13	2.71	2.37	2.40
9/8	3.2	84	10	1.79	2.24	2.28
9/9	2.8	88	11	2.40	2.39	2.44
9/10	2.2	92	8	3.26	2.45	2.52
9/11	3.6	100	9	2.71	2.96	3.00
9/12	4.0	98	13	2.64	2.98	3.00
9/13	3.4	95	-12	2.71	2.38	2.45
9/14	3.2	86	-8	2.40	2.03	2.10
9/15	2.6	84	2	2.30	2.05	2.11
9/16	2.5	88	15	2.08	2.42	2.48
9/17	2.4	92	19	2.56	2.65	2.71
9/18	2.7	82	18	2.48	2.23	2.27
9/19	2.6	79	17	2.20	2.08	2.12
9/20	2.5	80	19	2.08	2.14	2.18
9/21	2.3	78	20	2.40	2.05	2.10
9/22	2.8	74	23	2.08	1.98	2.00
9/23	2.4	80	13	2.48	2.03	2.09
9/24	2.3	80	17	2.40	2.09	2.14
9/25	3.8	80	12	2.20	2.17	2.18
9/26	2.6	75	-14	1.79	1.40	1.47
9/27	4.4	80	-6	2.49	1.94	1.95
9/28	3.9	74	-3	2.20	1.67	1.70
9/29	2.4	81	-7	2.30	1.75	1.83
9/30	2.2	91	15	2.71	2.52	2.59

3. The diurnal variation of NO₁, NO₂ and oxidant concentration.

The 10% trimmed mean was used here to study the diurnal patterns of NO₁, NO₂ and oxidant concentration at Redwood City and San Jose monitoring stations.

A 10% trimmed mean is the average of the remaining data after dropping the largest 10% of the data and the smallest 10% of the data. The influence function of 10% trimmed mean has two desirable properties--boundedness and continuity. But the influence function of the mean is unbounded and the influence function of the median is discontinuous. [4]

From the 10% trimmed means of NO hourly concentration for workdays (Figure 6), we see that an apparent difference between San Jose (SJ) and Redwood City (RC) occurs during the morning rush hour (6:00 a.m. - 7:00 a.m.) and the evening (5:00 p.m. - 12:00 p.m.) In the morning, NO in San Jose is higher than that in Redwood City. But in the afternoon, Redwood City had higher NO concentration than San Jose. This might indicate that in San Jose the traffic produces NO and the sunlight speeds the radiation reaction

where NO₂ was accumulated at night [5].

From the 10% trimmed means of NO₂ concentration (Figure 7), NO₂ at night is higher in San Jose than in Redwood City. Since the effect of afternoon high temperature will convert NO to NO₂ and the maximum temperature in San Jose usually is 15°F higher than in Redwood City, the concentration of NO in the evening is higher in Redwood City than in

San Jose. Also NO transported by wind from Redwood City to San Jose has already been converted to NO_2 . These might be the reasons why NO_2 in San Jose is always higher than that in Redwood City. From Figures 6 and 7 we also noticed that there is time delay for NO_2 to reach the highest point. For example, the peak hour for NO_2 in the morning is 9:00 a.m., but NO is 7:00 a.m.

Using the 10% trimmed means of oxidant concentration (Figure 8), we see that the diurnal variation in San Jose has the same pattern as that in Redwood City despite some meteorologic and geographic difference. Since more than 10% of the hourly oxidant concentration is below detection threshold (zero), then the 10% trimmed mean only affects the upper tail. Hence it seems reasonable to use the 10% trimmed mean to study the pattern of diurnal variation.

Figure 6

Figure 7

Figure 8

In Section 3, we used 10% trimmed means to study the diurnal variability of the pollutants. The standard mean and median are the common ways to investigate the diurnal pattern. The comparison among these three methods is shown in Figure 9 using hourly oxidant concentrations at San Jose.

Figure 9

REFERENCES

- [1] William S. Cleveland and Roberta Guarino, 1976. "Some Robust Statistical Procedures and Their Application to Air Pollution Data," Technometrics, Vol. 18, No. 4, November, 1976.
- [2] Bay Area Air Pollution Control District, Technical Services Division. Contaminant and Weather Summary, July, August, and September of 1974.
- [3] Albert E. Beaton and John W. Tukey, 1974. "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-spectroscopic Data," Technometrics, Vol. 16, No. 2, May, 1974.
- [4] Frank R. Hampel, 1974. "The Influence Curve and Its Role in Robust Estimation," Journal of the American Statistical Association, Vol. 69, pp. 383-393.
- [5] G. C. Tiao, M. S. Phadke and G. E. P. Box. "Some Empirical Models for the Los Angeles Photochemical Smog Data," Journal of the Air Pollution Control Association, Vol. 26, No. 5, May, 1975.