URBAN AIR POLLUTION AND MORTALITY: TEN YEARS OF PHILADELPHIA DATA

Ronald E. Wyzga,* Electric Power Research Institute

I. INTRODUCTION

In spite of the large amount of research that has been undertaken to learn about the relationship between human health and air pollution, we still remain relatively ignorant. We cannot enumerate all of the health impacts; we are not at all certain about the identity of those pollutants, singly or in combination, which may be responsible for health effects; and we are ignorant about the dose-response relationship between these two elements. Further information is badly needed about this relationship to guide us in the development of future energy technologies.

All believable energy scenarios for the U. S. indicate an important role for coal. Uncontrolled coal combustion produces significant air pollution. To date, we have regulated the combustion of coal and adopted technologies to reduce the emissions of particulates and SO_2. There are those who question whether present control technologies are sufficient or whether or not we should concentrate our pollution control efforts on other substances. This question is also posed by those developing new technologies in which some trade-offs may be necessary. For example, there are new technologies which could allow us to reduce our NO_x emissions further, but at the expense of increased polycyclic organic emissions. There is also concern that some currently suggested methods of SO_2 control in coal-fired power plants could lead to the increased formation of sulfate and other oxidized sulfur compounds.

This paper describes a model used to estimate the association between air pollution and health as measured by mortality and then tries to identify those pollutants which are more closely associated with mortality.

II. THE DATA AND VARIABLES

This study uses Philadelphia data for the years 1957-1966. Daily mortality data by cause of death are available for those residents of Philadelphia who died in that city. Two daily pollution measures were available for the ten-year period: coefficients of haze (smoke shade) and total suspended particulate (HI-VOL) measures. These two measurements were taken at two or three sampling stations in Philadelphia. Three mutually exclusive time periods (1957-1960, 1961-1963, and 1964-1966) are defined to accommodate changes in sampling sites over the ten-year period. Data from the same stations are then generally available for each day of a particular time period. This partition into three time periods allows three replications of each subsequent model examined and aids in the model and variable development. The first and third time periods of the study use data from two sampling stations. Coefficients of haze (COH) measurements and total suspended particulate (TSP) data from the two stations are generally available for each day of these time periods. Those days for which one or more station measures are missing are excluded from this study. For the second study period, data are available from three sampling stations. When daily measures were missing from one station, they were estimated through use of an iterative regression procedure (1). The COH variable for such a day would then be the average of the observed COH values and the estimated value. When measurements were missing for two or more stations on a given day in the second time period, that day was eliminated from the investigation.

For the third time period, measures of six additional pollutants are available from one monitoring station in Philadelphia. These pollutants are NO, NO_2, SO_2, hydrocarbon, CO and oxidants. The means and standard deviations of the pollution variables for the winter months of the 1964-1966 period are given in Table 1. Table 2 gives the estimated correlation coefficients for each pair of pollution variables during that period.

Several seasonality variables were compared and a weighted 30-day moving average of past temperatures, which gave twice the weight to the most recent 15 days, was chosen because it correlated more highly with total mortality than several other moving-averages of temperature and because it was significantly more highly correlated with mortality than Fourier functions of time.

The performance of the seasonal adjustment variable for the winter months differed significantly from its performance for the summer months. Accordingly, the year is divided into halves in subsequent analyses. An epidemic variable for the winter months and a heat-wave variable for the summer months were also found to be important contributors to the variation in mortality data. These variables are included in the following analyses. The epidemic variable is defined from the residuals obtained from regressing monthly New Jersey mortality data appropriately detrended on current and preceding Philadelphia temperature averages.

The heat-wave variable is the weighted product of lagged and unlagged values from a one-to-six corrected effective temperature scale. If the corrected effective temperature scale value for the day of mortality is represented by E(D), then the variable used can be written $E(D)^3 E(D-1)^2 E(D-2)$. For each season, a two-day moving-average of temperature which represents recent weather is also added to the set of variables included in the analysis.

The means and standard deviations of each variable analyzed for all time periods are given in Table 3.

III. DEVELOPMENT OF A MODEL

Regression models are considered. Total mortality was first regressed upon the group of adjustment variables and COH and TSP for the summers and winters of each of the three time periods. Given the high correlation between TSP and COH, it was felt that any further model development would best consider only one of the two variables, and since the regression coefficients of COH were associated with larger t-statistics than those of TSP, the COH variable was used in further model development.

Linear and non-linear models were considered, and linear models performed better than non-linear models and were therefore considered in the subsequent analyses. The residuals of the linear regression appeared to be normally distributed although they were serially correlated in some time periods.

IV. RESULTS

The results of the regressions on total mortality, using the COH (coefficient of haze) measure as a pollution variable, are given in Table 4. The COH coefficients are all positive, and those for all of the winter periods are for the 1957-1960 summer period are significant. The mean pollution levels for the two summer periods in which the COH coefficients are not significant are noticeably smaller than the pollution variables for the other time periods. This fact might explain the non-significance of the COH coefficients for these time periods.

There could be two reasons for detecting a weaker relationship between the COH values and mortality for those periods with smaller COH values. First of all, the response of mortality could be non-linear with a proportionately stronger response to higher pollution levels than a linear model suggests in spite of the fact that a linear model performed better than non-linear models tested. Functions using the COH values only above a certain threshold and exponential functions of the COH values were introduced into the regressions, but they gave no higher association with mortality than the initial COH variables. The second reason could explain a smaller association between the COH measure and mortality when the COH measures are small, even if the relationship were linear. The measurement error of smaller COH values is far greater relative to their size than the measurement error of the larger COH values. As measurement error would bias the regression coefficients of the COH variable downward (2,3), the coefficients of smaller COH variables would be subject to a greater bias than the coefficient of larger COH variables.

The significant Durbin-Watson statistics indicate the presence of serial correlation, which can lead to overestimates of the (absolute values of the) t-values used to test the significance of the coefficients (2,3). To adjust for this problem, a non-linear regression model incorporating serial correlation was fitted. The results showed no changes in the significance levels for any of the COH coefficients.

The beta coefficients (β coeff.) presented in Table 4 indicate the predicted number of standard deviations the mortality variable will change for each increase of one standard deviation in that variable, if one assumes that the other variables remain constant. Thus if the linear regression model is correct for the 1964-1966 winter data, the estimates predict that an increase of one standard deviation in the COH variable will lead to an increase of mortality on that day of 0.1349 times the standard deviation of total mortality (9.22) or to an increase of about one death.

The results indicate how important the epidemic and "heat wave" variables are in explaining daily mortality. The 2-day temperature variable in the summer months is also an important predictor of mortality, and it probably complements the "heat wave variable" as an index of hot weather.

Data for the other pollutants (NO, NO_2, SO_2, hydrocarbons, CO, and oxidant) were available only for the 1964-66 time period. Given the lack of significance of the COH variable coefficient in the 1964-66 summer period, the data for this period were not analyzed with the additional pollutants. The winter 1964-66 data were analyzed using a series of regression models similar to that described above, but with a different pollution variable in each regression. The series of regressions permitted a comparison of regression coefficients and avoided a multicollinearity problem which would have arisen given the degree of correlation between several pairs of pollutants. (See Table 2). Serial correlation was not statistically significant, and no adjustment was undertaken.

Table 5 summarizes the results of regressing the various pollution variables upon daily total mortality. The seasonality variable, two-day temperature variable and epidemic variable were also included as independent variables in these regressions. All of the pollution variables have positive coefficients, but only the COH, NO and hydrocarbon variables have significantly positive coefficients.

V. DISCUSSION

From the results it is difficult to generalize about which pollutant is best as an index, or which may affect health most. Differences in the estimated coefficients could be due to differences in measurement error among the pollutants. The greater the random measurement error of a variable, the larger the downward bias in the coefficient of that variable (2,3). As different measurement methods are involved in measuring the various pollutants, the measurement errors cannot be expected to be the same for each variable. Another source of error leading to the same type of downward bias is the local influence upon a variable. Local influence is the influence of nearby sources upon a pollution measure. These local influences can be thought to be a kind of measurement error imposed upon an overall urban index. The variables other than COH and TSP are particularly susceptible to this type of error, as measures from only one station are available.

The estimated increase in the number of deaths associated with an increase of one standard deviation in the pollution variable ranges from 1.24 when COH or NO are the pollutants in the regression to 0.25 when oxidant is the pollutant examined. These estimates only consider deaths on the day of pollution; lagged or delayed effects are not included here.

A model to examine lagged effects was developed (4) using the COH variable. The large number of missing observations for the other pollution variables made it difficult to apply lagged models with these variables. The model developed was a geometrically-distributed lag model which adjusted for serial correlation. This model was applied to the four time periods in which the COH variable was statistically significant and yielded similar estimates of the COH impact for each period. Table 6 presents the results of this model for the 1964-66 winter period. The total increase in the estimated impact of the pollution variable on mortality is about one third, with almost no impact of pollution occurring beyond two days after the pollution occurred.

Chronic or greatly delayed effects cannot be estimated with time series models of daily data.

VI. SUMMARY AND CONCLUSIONS

Environmental air pollution is associated with increased mortality. Although this association is significant, the other environmental phenomena, such as heat waves, may be responsible for a larger number of deaths.

The use of different pollution variables was investigated. One would expect the different pollution measures to perform quite similarly as meteorological conditions largely determine the concentrations of pollutants in the atmosphere. All of the pollutants were positively associated with mortality, but only variables derived from COH, NO and hydrocarbon measurements were significantly associated with mortality. Until further information is obtained about the effects of measurement error and local influence upon the various pollution measures, it is impossible to associate mortality more closely with one type of pollution than with another. It should also be added that it will be necessary to consider additional pollutants or combinations of pollutants. Certainly one hears the names of additional emitted compounds as one investigates new and existing energy technologies.

REFERENCES

1. Wyzga, R. E. (1973): "Note on a Method to Estimate Missing Air Pollution Data in A Statistical Analysis of the Relationship Between Daily Mortality and Air Pollution Levels". Journal of the Air Pollution Control Association, 23(3), 207-208.

2. Johnston, J. (1963): Econometric Methods. McGraw-Hill, New York.

3. Malinvaud, E. (1966): Statistical Methods of Econometrics. Rand-McNally and Co., Chicago.

4. Wyzga, R. E.: The Effect of Air Pollution upon Mortality: A Consideration of Distributed Lag Models. Submitted for publication.

TABLE 1

Means and standard deviations of pollution variables, 1964-1966

Variable			Winter Periods
COH[a]	(per 1000 ft.)	Mean	131.00
		S.D.	55.14
TSP	($\mu g/m^3$)	Mean	161.59
		S.D.	66.95
NO	(parts per hundred million)	Mean	6.36
		S.D.	5.59
NO_2	(parts per hundred million)	Mean	3.41
		S.D.	1.26
SO_2	(parts per hundred million)	Mean	9.57
		S.D.	6.90
Hydrocarbon	(parts per ten million)	Mean	22.65
		S.D.	7.02
CO	(parts per million)	Mean	7.54
		S.D.	3.11
Oxidant	(parts per hundred million)	Mean	2.02
		S.D.	1.24

[a] The COH variable has been multiplied by 100.

TABLE 2

Correlations between pollution variables, 1964-1966 winters

Variable	COH	TSP	NO	NO_2	SO_2	HC	CO	OX
COH	1.000	.795	.811	.600	.667	.688	.329	.403
TSP		1.000	.657	.629	.654	.570	.306	.302
NO			1.000	.596	.520	.625	.325	.435
NO_2				1.000	.544	.508	.203	.378
SO_2					1.000	.562	.074	.163
HC (Hydrocarbon)						1.000	.147	.458
CO							1.000	.233
OX (Oxidant)								1.000

TABLE 3

Means and standard deviations of variables

		Winters			Summers		
Variable		1957-60	1961-63	1964-66	1957-60	1961-63	1964-66
Total daily mortality	Mean	66.90	66.89	64.62	58.55	58.85	60.36
	S.D.	10.19	10.36	9.22	10.74	10.21	10.27
2-day moving-average temperature	Mean	90.16	87.18	89.56	149.79	149.66	151.15
	S.D.	25.94	27.25	22.32	19.51	18.81	19.98
30-day moving-average temperature	Mean	1008.49	978.03	1008.63	1692.83	1683.30	1680.62
	S.D.	200.15	241.08	181.40	148.81	148.33	191.23
COH variable[a]	Mean	189.42	160.84	131.00	122.02	92.39	87.13
	S.D.	76.28	69.23	55.14	44.11	37.62	41.33
Epidemic variable	Mean	21.09	19.95	19.20			
	S.D.	7.08	7.29	6.22			
Effective temperature function	Mean				502.75	531.33	533.47
	S.D.				1676.49	1691.38	1467.94

[a] The COH variable have been multiplied by 100

TABLE 4

Comparison of results from multiple regressions - total mortality

Variable		1957-1960	1961-1963	1964-1966
A. Winter Periods				
COH variable	Coeff	0.0098	0.0126	0.0226
	β Coeff.	0.0735	0.0840	0.1349
	t-value	2.00*	2.02*	2.85**
Seasonality variable	β Coeff.	0.3077	-0.4292	-0.2812
	t-value	-5.72**	-7.09**	-4.44**
2-day temperature variable	β Coeff.	-0.0165	0.0626	0.0742
	t-value	-0.31	1.12	1.18
Epidemic variable	β Coeff.	0.2000	0.3468	0.0820
	t-value	5.57**	9.40**	1.75
Multiple correlation coefficient squared (R^2)		0.1550	0.2879	0.0926
Durbin-Watson statistic		1.7622**	1.7652**	1.8809
Number of observations		660	532	421
B. Summer Periods				
COH variable	Coeff.	0.0286	0.0199	0.0052
	β Coeff.	0.1174	0.0734	0.0208
	t-value	3.53**	1.82	0.41
Seasonality variable	β Coeff.	0.4663	0.2407	0.4426
	t-value	-9.89**	-4.48**	-6.53**
2-day temperature variable	β Coeff.	0.2048	0.0794	0.2273
	t-value	4.11**	1.42	3.06**
Heat wave variable	β Coeff.	0.4269	0.4671	0.2730
	t-value	12.15**	10.83**	5.01**
Multiple correlation coefficient squared (R^2)		0.3194	0.2360	0.1522
Durbin-Watson statistic		1.6503**	1.7597**	1.3458**
Number of observations		688	540	386

*Significant at the 0.05 level.
**Significant at the 0.01 level.

TABLE 5

Comparison of results from multiple regressions, 1964-1966, total mortality upon various pollutants

Winter 1964-1966

Pollution Variable	β coefficient	t-value
COH	0.1349	2.85**
TSP	0.0808	1.85
NO	0.1347	3.11**
NO_2	0.0565	1.28
SO_2	0.0443	0.94
Hydrocarbon	0.0999	2.17*
CO	0.0620	1.35
Oxidant	0.0269	0.53

*Significant at the 0.05 level.
**Significant at the 0.01 level.

TABLE 6

Geometrically decreasing lag model with serial correlation 1964-1966 Winter Data

Parameter	Estimate	Std. Error of Estimate	t-value
30-day seasonality variable parameter	-0.0166	0.0038	-4.30**
2-day temperature variable coefficient	0.0349	0.0296	1.18
Epidemic variable coeff.	0.1637	0.0848	1.93
COH variable coeff. b	0.0204	0.0072	2.83**
Lag parameter λ	0.3251	0.0847	3.84**
Serial correlation ρ	-0.2671	0.0872	-3.06**
Total effect b/(1-λ)	0.0302	0.0106	2.85**

Regression Constant: 60.78

Source	Sum of Squares	Degrees of Freedom	Mean Square
Regression	4258.102	6	709.684
Residual	30035.984	390	77.015
TOTAL	34294.086	396	86.601

Multiple Correlation Coefficient R: 0.3524
R^2: 0.1242

* Significant at the 0.05 level.
** Significant at the 0.01 level.