00 Air Quality V12

Air quality and acute deaths in California, 2000-2012

S. Stanley Young, CG Stat, Raleigh, NC.
Kenneth K. Lopiano, Winterville, NC
Richard L. Smith, University of North Carolina, Chapel Hill, NC

Abstract

Many studies have sought to determine whether there is an association between air quality and acute deaths in the US. Additionally, many consider it plausible that current levels of air quality cause acute deaths. However, several factors call causation and even association into question. Multiple testing and multiple modeling and various biases can lead to false positive findings.

Moreover, the fact that most data sets used in studies evaluating the relationships among air quality and public health outcomes are not publicly available makes reproducing the results nearly impossible. Here we have publicly available a dataset containing daily air quality levels, $PM_{2.5}$ and ozone, daily temperature levels, minimum and maximum and daily relative humidity levels for the eight most populous California air basins. Over two million death certificates were obtained from the state of California and daily death counts in the eight air basins were derived. We analyzed the dataset using a standard time series analysis, a moving median analysis, and a prediction analysis in which we use leave-one-year-out cross validation analysis to evaluate predictions. Both standard time series analysis and the moving medians analysis found little evidence for association between air quality and acute deaths. The prediction analysis process was a run as a large factorial design using different models. We use holdout predictive mean square error to assess prediction. Among the variables used to predict acute death, most of the daily death variability was explained by time of year or weather variables. In summary, neither $PM_{2.5}$ nor ozone added appreciably to the prediction of daily deaths. The empirical evidence is that current levels of air quality, ozone and $PM_{2.5}$, are not causally related to acute deaths for California.

some of the coefficients $\beta_2,\ldots,\beta_k$ are constrained to be equal; however, this usually has only minor impact on the important coefficient β_1). Other common approaches use any of lags 0, 1, 2 in a single-lag model, or averages over any combination of lags 0, 1, 2, 3. For the present study, we have tried different combinations of lags to look for the lag combination that best represents the air pollution effect. We believe this approach to be justified in view of the weak evidence for any air pollution effect in these dataset; however, in view of the selection bias inherent in such an approach, we caution against over-interpretation of such results, especially in cases where the p-value is over 0.01 or the result highly depend on the selection of a particular combination of lags.

South Coast Air Basin

The approach outlined in the previous section is applied to data from each of eight California air basins, **Figure. 1.** Because they are the two most populated air basins, we concentrate initially on the South Coast air basin (which includes Los Angeles, Orange, Riverside and San Bernardino Counties) and the San Francisco Bay air basin (San Francisco, Marin, Sonoma, Napa, Solano, Contra Costa, Alameda, Santa Clara and San Mateo counties). For the response variable in this analysis, we use total non-accidental mortality among people aged 65 and over.

Fitting the meteorological model alone, in **Table 1** we tabulate the p-value associated with dropping each of the six terms in turn. Five of the meteorological variables are very highly significant; the only exception is current-day relative humidity. This result is based on the particular choices $df_0=7$, $df_1=df_2=6$, but the overall conclusion is robust against alternative values of those three degree of freedom parameters.

Variable	Lags	p-value
Daily Max Temperature	Current day 0	<1 e-16
Daily Max Temperature	Mean of 1,2,3	4.6 e-7
Daily Min Temperature	Current day 0	2.5 e-4
Daily Min Temperature	Mean of 1,2,3	2.4 e-5
Mean Daily Relative Humidity	Current day 0	0.18
Mean Daily Relative Humidity	Mean of 1,2,3	1.5 e-10

Table. 1: Statistical significance of meteorological components: based on model (1) without air pollution component and with $df_0=7$, $df_1=df_2=6$, fitted to nonaccidental mortality for ages 65 and up, South Coast air basin.

In subsequent analyses, we have retained all six meteorology components; this is to ensure consistency across different air basins and to avoid the analysis being biased by overuse of statistical significance tests; however, **Table. 1** is evidence that we have identified appropriate meteorological variables for the overall analysis.

We now consider addition air pollution variables to the meteorological model in **Table 1**. Initially, we concentrate on ozone. **Table 2** shows the coefficient estimates, standard error (SE), t-value and p-value associated with ozone at various combination of lags. The units here are percent rise in mortality per 10 ppb rise in ozone. The strongest positive coefficient is based on lags 0, 1, 2 and 3, for which the model predicts a 0.1% rise in mortality per 10 ppb rise in ozone. However, neither this nor any of the other values in the table comes anywhere close to being statistically significant. This is for 13 years of data over one of the most densely populated areas of the US – if there is an ozone-mortality effect in California, we ought to see it here.

Lags Included	Estimate	SE	t-value	p-value
0	0.0870	0.1135	0.77	0.44
1	-0.0472	0.1136	-0.42	0.68
2	0.0471	0.1141	0.41	0.68
0,1	0.0266	0.1315	0.20	0.84
1,2	0.0002	0.1330	0.00	1.00
0,1,2	0.0825	0.1507	0.55	0.58
0,1,2,3	0.1222	0.1673	0.73	0.46
0,1,2,3,4	0.0941	0.1802	0.52	0.60
0,1,2,3,4,5	0.0096	0.1905	0.05	0.96
0,1,2,3,4,5,6	-0.0479	0.1992	-0.24	0.81

Table. 2: Statistical significance of ozone component with various combinations of lags: based on model (1) $df_0=7$, $df_1=df_2=6$. Estimate is percent rise in mortality for 10 ppb rise in ozone. South Coast air basin; response variable is non-accidental mortality aged 65 and over.

The same analysis was tried using $PM_{2.5}$ in place of ozone, with results shown in **Table 3.** In this case, several of the estimates appear to be statistically significant with a p-value <0.05 (smallest value 0.017), but all the statistically significant values are negative, which is not biologically plausible. We conclude that either the small p-values are an artifact of the selection effect already mentioned, or there is some other biological mechanism, such as confounding by some other pollutant, that explains these results.

Lags Included	Estimate	SE	t-value	p-value
0	0.1261	0.0998	1.26	0.21
1	-0.1966	0.0990	-1.99	0.05
2	-0.2121	0.0995	-2.13	0.03
0,1	-0.0425	0.1144	-0.37	0.71
1,2	-0.2720	0.1151	-2.36	0.018
0,1,2	-0.1133	0.1294	-0.88	0.38
0,1,2,3	-0.1636	0.1409	-1.16	0.25
0,1,2,3,4	-0.1611	0.1499	-1.07	0.28
0,1,2,3,4,5	-0.2609	0.1582	-1.65	0.10
0,1,2,3,4,5,6	-0.2435	0.1659	-1.47	0.14

Table 3: Statistical significance of $PM_{2.5}$ component with various combinations of lags: based on model (1) $df_0=7$, $df_1=df_2=6$. Estimate is percent rise in mortality for 10 $\mu g/m^3$ rise in $PM_{2.5}$. South Coast air basin; response variable is non-accidental mortality aged 65 and over.

In these analyses, the overdispersion parameter was of the order of 1.07 – in other words, the variance of the mortality variables is inflated by a factor of 1.07 compared with the Poisson distribution. This is typical for this kind of analysis and does not indicate a problem. A much larger overdispersion parameter could indicate some important missing covariates.

San Francisco Bay Air Basin

So far, we have only considered one air basin. The second most populated is San Francisco Bay, which has substantially different weather patterns and demographics from the Los Angeles area. Therefore, the entire analysis has been repeated for this air basin, as a test of how robust the analyses are for different regions of the state.

Table 4 shows the statistical significance of the individual meteorology components, analogous to **Table 1** for the South Coast air basin. The main difference from **Table 1** is that neither of the components due to relative humidity is statistically significant. (Although not reported in the table, if both relative humidity components – current day and the average of lags 1, 2, 3 – are dropped together, rather than one at a time, we also do not get a statistically significant component due to relative humidity.) In the following analyses, to maintain consistency of analysis methods across different air basins, the main results are still reported including relative humidity, but to assess the sensitivity to this component, some of the analyses have been repeated omitting relative humidity altogether.

Variable	Lags	p-value
Daily Max Temperature	Current day 0	9.05E-11
Daily Max Temperature	Mean of 1,2,3	0.0071
Daily Min Temperature	Current day 0	0.0019
Daily Min Temperature	Mean of 1,2,3	0.043
Mean Daily Relative Humidity	Current day 0	0.41
Mean Daily Relative Humidity	Mean of 1,2,3	0.32

Table 4: Statistical significance of meteorological components: based on model (1) without air pollution component and with $df_0=7$, $df_1=df_2=6$, fitted to nonaccidental mortality for ages 65 and up, San Francisco Bay air basin.

Table 5 shows the results when ozone is added to the analysis. As with our earlier analyses for the South Coast air basin, none of the estimates of the ozone effect at various lags is statistically significant at the 0.05 level. However, two of the analyses (with lag 0 alone, and with lags 0 and 1 together) are statistically significant with a p-value of about .02 if the relative humidity component is omitted. This result illustrates the principle that if enough different models are tried, it is usually possible to find some model that gives a statistically significant result: it does not imply that the result is significant in any practical sense. It should also be noted, however, that all the coefficients of models that include lag 0 are similar in magnitude (between 0.3 and 0.6): the variation in p-values is mostly due to their standard errors.

Lags Included	RH included?	Estimate	SE	t-value	p-value
0	yes	0.4464	0.2471	1.81	0.071
1	yes	0.1889	0.2413	0.78	0.43
2	yes	-0.1560	0.2442	-0.4	0.52
0,1	yes	0.4909	0.3030	1.62	0.11
1,2	Yes	0.0225	0.2947	0.08	0.94
0,1,2	Yes	0.3281	0.3502	0.94	0.35
0,1,2,3	Yes	0.4210	0.3927	1.07	0.28
0,1,2,3,4	Yes	0.4716	0.4167	1.13	0.26
0,1,2,3,4,5	Yes	0.4703	0.4310	1.09	0.28
0,1,2,3,4,5,6	Yes	0.3325	0.4448	0.75	0.45
0	No	0.4838	0.2121	2.28	0.023
0,1	No	0.5948	0.2604	2.28	0.022

Table 5: Statistical significance of ozone component with various combinations of lags: based on model (1) $df_0=7$, $df_1=df_2=6$. Relative humidity is omitted from some of the analyses. Estimate is percent rise in mortality for 10 ppb rise in ozone. San Francisco Bay air basin; response variable is non-accidental mortality aged 65 and over.

Table 6 shows the corresponding results for $PM_{2.5}$, where again relative humidity has been omitted from some of the analyses to illustrate the sensitivity to this component. Our conclusions are similar: some rows of this table show a statistically significant effect with a p-value of the order 0.02, but taking account of the number of models examined in order to achieve this result, it is unlikely to be of practical significance.

The overdispersion parameter for these analyses was around 1.05.

Lags Included	RH included?	Estimate	SE	t-value	p-value
0	Yes	0.3031	0.2362	1.28	0.20
1	Yes	0.1235	0.2373	0.52	0.60
2	Yes	0.3769	0.2312	1.63	0.10
0,1	Yes	0.3968	0.2700	1.47	0.14
1,2	Yes	0.4614	0.2679	1.72	0.09
0,1,2	Yes	0.5903	0.3067	1.92	0.05
0,1,2,3	Yes	0.5688	0.3297	1.72	0.08
0,1,2,3,4	Yes	0.5042	0.3482	1.45	0.15
0,1,2,3,4,5	Yes	0.5500	0.3634	1.51	0.13
0,1,2,3,4,5,6	Yes	0.4884	0.3767	1.30	0.19
0,1,2,3	No	0.5712	0.3123	1.83	0.07
0,1,2,3,4	No	0.6518	0.3341	1.95	0.05
0,1,2,3,4,5	No	0.8169	0.3535	2.31	0.021
0,1,2,3,4,5,6	No	0.7737	0.3702	2.09	0.037

Table 6: Statistical significance of $PM_{2.5}$ component with various combinations of lags: based on model (1) $df_0=7$, $df_1=df_2=6$. Relative humidity is omitted from some of the analyses. Estimate is percent rise in mortality for 10 $\mu g/m^3$ rise in $PM_{2.5}$. San Francisco Bay air basin; response variable is non-accidental mortality aged 65 and over.

Combining Results Across Air Basins

In the NMMAPS papers on ozone **[9, 18]**, the single-city analyses were repeated for up to 98 US cities for which ozone and mortality data were available. They were then combined across cities using a hierarchical model analysis, based on an algorithm originally due to Everson and Morris **[21]** and coded by Roger Peng into the R function "tlnise" **[22]**. The same method is used here to produce estimates that are combined across all eight air basins in our study. It would not be practicable (or interpretable) to repeat all the analyses for every combination of meteorological variables, lags of the pollutant variable, or degrees of freedom for the spline components of the model. Therefore, some choices were made, guided by the analyses already conducted for the South Coast and San Francisco Bay air basins, as follows:

1. All analyses used all six meteorological variables.
2. The degree of freedom parameters were set to be respectively 7, 6 and 6, for df_0, df_1 and df_2.
3. For both ozone and $PM_{2.5}$, only certain combinations of lags were tried.

The results of this analysis are shown in **Table 7.** None of the analyses show a statistically significant effect when combined across all eight air basins.

Variable	Lags	Estimate	SE	t-value	p-value
Ozone	0,1	0.3376	0.2434	1.39	0.17
Ozone	0,1,2	0.3165	0.2466	1.28	0.20
Ozone	0,1,2,3	0.4149	0.3260	1.28	0.20
PM2.5	0,1	0.0126	0.2034	0.06	0.95
PM2.5	0,1,2,3	-0.0006	0.2464	0.00	1.00
PM2.5	0,1,2,3,4,5	0.0689	0.2799	0.25	0.81

Table 7: Combined results across all eight air basins.

All the analyses in this paper so far are based on total non-accidental mortality for ages 65 and up. The analysis was repeated using (a) total non-accidental mortality for all ages, (b) respiratory deaths aged 65 and up, (c) circulatory deaths aged 65 and up, (b) combined respiratory and circulatory deaths aged 65 and up. None of these produced a statistically significant result in the combined analyses.

The results of **Table 7** were also repeated with the choices df_0=7, df_1=6, df_2=6 replaced by (a) df_0=10, df_1=6, df_2=6, (b) df_0=7, df_1=3, df_2=3, (c) df_0=10, df_1=3, df_2=3. The analysis of **Table 7** was also repeated with relative humidity omitted from the analysis. None of these changes produced a statistically significant result in any of the combined analyses.

[Point to Sup 3.2 A and Sup 3.2 B here]

Comparisons with NMMAPS

We have pointed out that the statistical methods of this paper are similar to those of the NMMAPS study; see in particular **[8, 15]**, but they are not identical. Those papers also included an interaction effect between age and long-term trend, and the meteorological variables were daily mean temperature and dewpoint, rather than those of the present paper. What happens if we use exactly the same methods for the two datasets?

To investigate this question, we recompiled the NMMAPS dataset but using tmax, tmin and daily max relative humidity as the meteorological variables. (Those variables are all in the NMMAPS dataset, but were not used in the previously cited papers.) The dataset was analyzed using the same computer code as the other analyses in this paper, applied to deaths aged 65 and over analyzed as a single age group (no interactions). We took $df_0=7$, $df_1=df_2=6$ as in most of the analyses in this paper, and the distributed lag structure based on lags 0 through 6.

Since the rest of this paper is concerned with California data, we concentrated on the California cities in the NMMAPS database. **Table 8** shows results for each city, and the combined result for all 12 California cities. Also shown in **Table 8** is the national result, in which the 12 California cities were combined with 86 other US cities, reanalyzed using the software of the present paper.

City	Estimate	SE	t-value	p-value
Bakersfield	0.7031	0.9970	0.71	0.48
Fresno	0.1577	0.9520	0.17	0.87
Los Angeles	0.1941	0.2199	0.88	0.38
Modesto	0.3027	1.5057	0.20	0.84
Oakland	0.8943	1.0210	0.88	0.38
Riverside	0.0255	0.6019	0.04	0.97
Sacramento	-0.0913	0.8334	-0.11	0.91
San Bernardino	0.7358	0.6330	1.16	0.25
San Diego	0.1080	0.4717	0.23	0.82
San Jose	-0.0481	0.9756	-0.05	0.96
Santa Ana Anaheim	0.1231	0.4815	0.26	0.80
Stockton	0.9981	1.3775	0.72	0.47
All CA	0.2485	0.2307	1.08	0.28
National	0.2873	0.0915	3.14	0.0017

Table 8: Estimates for the ozone effect in 12 California cities from the NMMAPS study (San Francisco omitted because of lack of ozone data). Also shown are the combined results from all 12 cities under "All CA", and the combined results of all 98 US cities included in the NMMAPS ozone study. Applied to all deaths aged 65 and up, using tmax, tmin and maximum relative humidity as the three meteorological variables, and a distributed lag model for ozone covering lags 0-6.

The last result shows a combined estimate of 0.287 (percent rise in mortality per 10 ppb rise in 8-hour daily max ozone) and a standard error (more precisely, posterior standard deviation) of 0.0915. By comparison, the result quoted in Smith **[9]** was a combined estimate of 0.411 and a posterior standard deviation of 0.080. Just to make a further comparison with the results of Smith **[9],** the method of the present paper was repeated with mortality data from all age groups 55 and up (the same as in the original NMMAPS analyses) – in this case our estimated combined national coefficient, using the meteorological model of the present paper, rises only very slightly, from 0.287 to 0.300. Therefore, the difference in combined estimates compared with Smith **[9]** appears to be due to the different meteorological variables used and not to the different