A two-stage model for multiple time series data of counts

KIROS BERHANE*, DUNCAN C. THOMAS
Department of Preventive Medicine, University of Southern California, Los Angeles, CA 90033-9987, USA
kiros@rcf.usc.edu

SUMMARY
We propose a two-stage model for time series data of counts from multiple locations. This method fits first-stage model(s) using the technique of iteratively weighted filtered least squares (IWFLS) to obtain location-specific intercepts and slopes, with possible lagged effects via polynomial distributed lag modeling. These slopes and/or intercepts are then taken to a second-stage mixed-effects meta-regression model in order to stabilize results from various locations. The representation of the models from the stages into a combined mixed-effects model, issues of inference and choices of the parameters in modeling the lag structure are discussed. We illustrate this proposed model via detailed analysis on the effect of air pollution on school absenteeism based on data from the Southern California Children’s Health Study.

Keywords: Air pollution; Correlated data; GEE; Mixed effects; Poisson regression; Time series.

1. INTRODUCTION

Time series methods have been widely used to investigate the ‘acute’ effects of daily fluctuations in ambient air pollution on the corresponding rates of various health events, including mortality (Schwartz, 2000b), hospitalization (Schwartz, 1995), and school absenteeism (Ransom and Pope, 1992). The majority of these studies were conducted in single cities, although recently Dominici et al. (2000) have described analysis of mortality data from several US cities. Here, we describe some new approaches to such studies, as illustrated by a study of school absence rates in 11 Southern California communities (Gilliland et al., 2001). This study was part of a larger project conducted in a cohort of fourth, seventh and tenth graders enrolled in the fall of 1992 and a second cohort of fourth graders enrolled in the fall of 1995 (Peters et al., 1999). The larger study was aimed at investigating the ‘chronic’ effects of four air pollutants: namely, PM$_{10}$, O$_3$, NO$_2$ and inorganic acid (with 90% nitric acid and 10% hydrochloric acid).

In the school absenteeism substudy, all absences in the second cohort of 2081 fourth graders during the period from January to June of 1996 were ascertained and telephone interviews were conducted to determine whether the absences were illness related and to collect more specific information on symptoms, physician visits, and previous activities. Details of the epidemiologic methods and results are provided in Gilliland et al. (2001).

In estimating the effects of air pollution on school absenteeism in a time series of rates, several key characteristics of the data must be taken into account. First, the expected counts must be constrained to be positive, suggesting a log-linear regression. The denominators of the rates enter the log-linear model

*To whom correspondence should be addressed

© Oxford University Press (2002)
on the counts as an offset in order to take into account the varying number of ‘at risk’ subjects. Second, the data may exhibit extra-Poisson variation due to the effect of unmeasured factors on the daily counts and cannot be directly controlled for in regression models (McCullagh and Nelder, 1989). An appropriate statistical analysis should also account for the serial correlation in the data. Finally, the daily outcome of counts are not only the function of the same-day exposures to pollutant(s), but are more likely to be affected by exposures from a certain lag period. The polynomial distributed lag model of Almon (1965) has been used to account for the lag structure in epidemiologic analyses of time series data (Schwartz, 2000a).

For a single time series of counts, a method known as the ‘iteratively weighted filtered least squares’ (IWFLS) has been developed (Zeger, 1988; Samet et al., 1995). A related alternative model (Kelsall et al., 1999) uses a generalized additive model with an additional smooth term for time in order to deal with the long-term time trend and autocorrelation in the data. In our absence monitoring example, the 12 communities differ by design in their ambient pollution levels and hence averaging over these levels is not meaningful since we are interested in making both within-community and between-community comparisons. In this paper, we develop an analytic framework that allows for multi-stage modeling while properly handling the four key features of each time series of counts from many locations, hereafter referred to as communities. We propose a two-stage model. In the first stage, we fit a modified IWFLS model that allows for within-community comparison and produces community-specific intercepts along with their variance estimates and parameters that account for community-specific terms of polynomial distributed lag effects. The community-specific lag effects (and their variance estimates) are then recovered from the distributed lag terms. In the second stage, linear ‘ecologic’ models are fitted that summarize the combined ‘acute’ effects of pollution. Models that have at least two stages have a long history in statistics dating back to Rowell and Walters (1976).

We also propose linear mixed-effects second-stage models that allow for tests of an overall summary of the ‘acute’ effects and possible heterogeneity between them for each set of lagged estimates, or an appropriate summary thereof, in the style of Stram (1996). The community-specific intercepts may also be fitted against the ambient pollutant levels via an ‘ecologic’ linear regression, weighted by the inverse of the variances, to provide ‘chronic’ effect estimates. The proposed approach differs from those proposed by Dominici et al. (2000) and Schwartz (2000a) in a number of ways. First, our approach models the autocorrelation in the time series directly via separate regression models in contrast to using a smooth term on time to account for the autocorrelation in the data. Second, our approach does not use Bayesian techniques to pool evidence from multiple communities.

The remainder of the paper will be organized as follows. In Section 2, we describe our two-stage technique for multiple Poisson time series data. In Section 3, we provide a detailed description and analysis of the absence monitoring data from the CARB-USC Children’s Health Study. Conclusions and future directions are discussed in Section 4.

2. THE PROPOSED TWO-STAGE MODEL

To fix ideas, we first discuss details about the models for a single time series of counts. Samet et al. (1995) proposed an extension of Poisson regression (McCullagh and Nelder, 1989) to account for key characteristics time series data of counts. This approach, known as IWFLS, assumes that the logarithm of the expected count is a linear function of the predictor variables. To account for overdispersion, it assumes that the variance is proportional to the mean $\mu_{ct}$, with proportionality constant, $\phi_{ct}$, that is assumed to be a smooth function of time, $t$, the mean, $\mu_{ct}$, and their product, $\mu_{ct} \times t$. The overdispersion term $\phi_{ct}$ is estimated by regressing the squared, standardized residuals from the log-linear model $t$, $\mu_{ct}$ and $\mu_{ct} \times t$ using a generalized additive model (Hastie and Tibshirani, 1990). The interaction term, $\mu_{ct} \times t$, is designed
to account for any complex nonlinear relationship in the effects of \( t \) and \( \mu_{ct} \) on overdispersion. This approach is similar to that of Zeger (1988), but both the overdispersion and autoregressive regressions are included in our model to more effectively weight the data in the log-linear analysis. For data with strong non-stationary correlation structure, one could use ‘empirical’ variance estimates as in Lumley and Heagery (1999). In the first stage, we propose to use an IWFLS model that properly analyzes the multiple time series of counts arising from the multiple communities. We refer the reader to Schwartz (2000a) and Dominici et al. (2000) for details on alternative generalized additive modeling approaches. Next, we give some details on the estimation process in the IWFLS approach.

### 2.1 First-stage models

Let \( Y_{ct} \) denote the daily count of incident absences in community \( c \) on day \( t \), and let \( N_{ct} \) denote the number of children ‘at risk’ for absenteeism on that day. A IWFLS-based Poisson regression model that has community-specific slopes and intercepts, with adjustment for other covariates \( W_{ct} \), is fitted in the first stage by organizing the correlation, and hence the variance–covariance, structures in a block-diagonal setup, under the assumption of independence between communities. This model is of the form

\[
\mu_{ct} = E(Y_{ct}) = N_{ct} \exp[a_{c} + s(t) + W_{ct}^{T} \zeta + b_{c}(Z_{ct} - \bar{Z}_{c})]
\]

leading to a log-linear model with \( \log(N_{ct}) \) as an offset. Note also that the \( s(t) \) term in (1) is only intended to account for the long-term time trends and usually uses 2–3 degrees of freedom. For details on the original IWFLS algorithm, we refer the reader to Zeger (1988) and Samet et al. (1995). More details on the modifications to the IWFLS algorithm needed for the multiple time series setup can also be found in Technical Report #152 from the authors at \texttt{http://www.hydra.usc.edu/biostat}.

The main steps for estimating the complete vector of regression parameters, say \( \beta \), can be summarized as follows:

1. Use the weighted and filtered least-squares technique to obtain updated estimates of \( \hat{\beta} \).
2. For each community \( c = 1, \ldots, C \), regress the squared residuals, \( r_{ct}^{2} \), on the fitted values, \( \hat{\mu}_{ct} \), time \( t \), and \( \hat{\mu}_{ct} \times t \) using gamma generalized additive model with log link function to obtain updated time-dependent overdispersion estimates, \( \hat{\phi}_{ct} \). Alternatively, one could also estimate a constant overdispersion factor, \( \hat{\phi}_{c} \).
3. For each community, calculate \( r_{ct}^{2} = (y_{ct} - \hat{\mu}_{ct})/\sqrt{\hat{\phi}_{ct} \hat{\mu}_{ct}} \) and obtain updated autoregressive parameter estimates, \( \hat{\alpha}_{c1}, \ldots, \hat{\alpha}_{cq} \), by regressing \( r_{ct}^{2} \) on \( r_{c,t-1}^{2}, \ldots, r_{c,t-q}^{2} \).
4. Iterate between the above three steps until convergence.

We start the algorithm with the Poisson regression coefficients \( \hat{\beta}, \hat{\phi}_{ct} = 1 \), and \( \hat{\alpha}_{c1} = \cdots = \hat{\alpha}_{cq} = 0 \). The simplification in computing gained from approximating the true variance, \( \hat{\Sigma} \), by the ‘model-based’ estimates, \( \hat{\Sigma} \), comes with penalty. It has been shown (Samet et al., 1995) that the model-based variance–covariance matrix is efficient only when dealing with low and stationary autocorrelation. Its asymptotic distribution is Gaussian with mean \( \beta \) but the model-based variance is inconsistent (Liang and Zeger, 1986). An ‘empirical’ variance estimate, which is consistent even when the correlation structure is misspecified, may be obtained by using the weighted empirical adaptive variance estimator (WEAVE) of Lumley and Heagery (1999).

When the residuals are not independent between communities, one needs to incorporate the spatial correlation into the modeling process. Dominici et al. (2000) discuss Bayesian hierarchical modeling of such data in detail. While this is conceptually simple to incorporate in our proposed methodology, it leads to a more complicated correlation structure. Thus, the computational advantages of the IWFLS approach may be lost since there may not be an obvious filtering process. We do not pursue this option any further in this paper.
An unconstrained distributed lag model corresponding to (1) would be

$$\mu_{ct} = E(Y_{ct}) = N_{ct} \exp\left[ a_c + s(t) + W^T_c \zeta + \sum_{l=0}^{L} b_{cl}(Z_{c,t-l} - \bar{Z}_c) \right].$$

(2)

However, the lagged terms are likely to be highly collinear due to autocorrelation in the exposure series. One solution, originally proposed by Almon (1965) and extensively used in the econometrics literature, is the use of polynomial distributed lag models. In this model, the effects of the lagged exposures are assumed to lie on a polynomial of sufficient degree $D$, i.e., $b_{cl} = \sum_{d=0}^{D} g_{cd} l^d$, so model (2) becomes

$$\mu_{ct} = E(Y_{ct}) = N_{ct} \exp\left[ a_c + s(t) + W^T_c \zeta + \sum_{d=0}^{D} g_{cd} \sum_{l=0}^{L} (Z_{c,t-l} - \bar{Z}_c) l^d \right].$$

(3)

Letting $M_c$ denote a matrix with $(l,d)$th element given by $l^d$, $g_c = (g_{c0}, \ldots, g_{cD})$ and $b_c = (b_{c0}, \ldots, b_{cL})$, the lag-specific ‘acute’ effect of pollution for community $c$ is given by

$$b_c = M_c g_c.$$

Rearranging model (3) by letting $\tilde{Z}_c = (Z_{c,t-0} - \bar{Z}_c, \ldots, Z_{c,t-L} - \bar{Z}_c)$, we get

$$\mu_{ct} = E(Y_{ct}) = N_{ct} \exp[a_c + s(t) + W^T_c \zeta + \tilde{Z}_c^T M_c g_c].$$

(4)

from which estimates for the lag-specific parameter vector, $b_c = M_c g_c$, could be obtained. Thus, the estimates for $b_c$ and its corresponding variance can be recovered from those for $g_c$ and $\text{var}(g_c)$, respectively. A test for a linear combination of the lagged ‘acute’ pollution effects can be conducted via a linear contrast such as $b_c = J b_c$. For example, $J = (1, \ldots, 1)$ gives an aggregate estimate of the overall ‘acute’ effect of pollution, summed over the entire lag period, under the polynomial distributed lag structure.

In models that incorporate polynomial distributed lag terms, we need to choose the number of lag days ($L$) and the degree of the polynomial ($D$). These choices are usually guided by the length of the time series and the substantive considerations. In our analyses, we search over a grid of $D$ and $L$ values and choose the ones that minimize the Akaike information criterion (Akaike, 1973) of the first-stage model(s). An alternative approach is to choose the final distributed lag structure based on second-stage results and/or a combination of results from both stages. One could also use the polynomial distributed lag modeling process to identify the most appropriate lag period. Then, all inference could be based on a summary index over the appropriate lag period. This could potentially lead to simple-to-understand and easily interpretable results.

2.2 Second-stage models

The first-stage regression (1) is followed by two ecologic linear regression models given by

$$a_c = \omega_0 + \omega (\bar{Z}_c - \bar{Z}) + \epsilon_c$$

(5)

$$b_c = \delta_0 + \delta (\bar{Z}_c - \bar{Z}) + \eta_c.$$  

(6)

Here, $\omega$ (from (5)) is the slope of the regression of community-averaged log absence rates on average pollutant levels, and hence provides ‘ecologic’ estimates of ‘chronic’ effects of pollution. On the other hand, $\delta_0$ from (6) is the mean of the within-community slopes, $b_c$, and hence serves as an aggregated
A two-stage model for multiple time series data of counts

Acute-effect estimate. Here, $\delta_0$ is a modified form of the community-specific slopes by average pollution levels and is the quantity of primary interest. Note that we use the deviation of the daily exposure values from $\bar{Z}_c$, which may be the pollutant of interest from stage 1 or any other ecologic covariate, in order to make the within-community and between-community comparisons of pollution effects orthogonal to each other. The parameter $\delta$ characterizes the relationship between the six month average pollution levels and the daily within-community pollution exposure levels. The stage two regressions are weighted by the inverses of the variances of $a_c$ and $b_c$, respectively.

For model (3), we propose a second-stage mixed-effects model that allows statistical inferences on the aggregated ‘acute’ effects of pollution in the style of Ware and Stram (1988) and Stram (1996). The model is

$$b_{cl} = \delta_0 + \delta(Z_c - \bar{Z}) + \eta_{cl} + \epsilon_{cl}$$  \hspace{1cm} (7)

where $\eta_{cl}$ is a random effect that has a known variance–covariance matrix given by the variance of $\hat{b}_{cl}$ from stage one, discussed in Section 2.1. We assume that $\eta \sim MVN(0, \Sigma)$ and $\epsilon \sim MVN(0, \tau^2I)$. This allows us to appropriately weight the regression by the block-diagonal matrix $V$ and also adequately conducts the meta-regression that summarizes the ‘acute’ effects of pollution across communities. Given an overall design matrix $Z$ and letting $\delta = (\delta_0, \delta_1)$, where $\delta_0 = (\delta_{01}, \ldots, \delta_{0L})$, one could conduct inference on the overall aggregated effect of pollution over the entire lag period, i.e.

$$H_0 : \mathcal{J} \delta_0 = 0$$

where $\mathcal{J}$ is a $L \times 1$ vector of 1’s, $\text{Var}(\hat{\delta}) = (Z^T \hat{\Sigma}^{-1} Z)^{-1}$ and $\hat{\Sigma} = \hat{\epsilon}^2 I + V$. An estimate of $\tau^2$ is obtained from the data via an iterative process. For more details on the estimation procedure, see Stram (1996).

The models in the two stages could be combined to form a generalized linear mixed model. For example, combining (4), (5) and (7) leads to

$$\mu_{ct} = E(Y_{ct}) = N_{ct} \exp[\alpha_0 + a \bar{Z}_c + s(t) + W^T_\eta \xi + \bar{Z}_c^T \delta_0 + Z_c \delta + \epsilon + \eta_c \hat{Z}_c]$$  \hspace{1cm} (8)

where $\epsilon$ and $\eta_c$ are Gaussian error terms with means zero and variances $\sigma^2_\epsilon$ and $\sigma^2_{\eta_c}$, respectively. This setup allows us to conduct the estimation process in a unified manner. Then, a test on the overall ‘acute’ effect of pollution, aggregated over the entire lag period, would take the form $H_0 : \mathcal{J} \delta_0 = 0$.

Operationally, the use of existing software for generalized mixed-effects models may not be computationally efficient. The community-specific time series could be very long (possibly involving daily time series from several years) and inverting the correlation matrices directly may not be practical. Hence, we prefer to use the two-stage approach where the computational advantages of the IWFLS approach can be exploited. However, we also point out that the combined model would provide more stable second-stage estimates, especially when dealing with distributed lag estimation techniques.

3. THE ABSENCE MONITORING EXAMPLE

In this section, we present a detailed analysis of the absence monitoring data set to examine the relationship between air pollution and school absenteeism. Absence reports were obtained from the schools these children were attending, generally within four weeks after the first day of absence. Trained interviewers telephoned the parents after each absence to inquire about the reason for the absence (whether illness-related and if so, the symptoms) and activities immediately preceding the illness, provided that the requirement for a four week reporting time window was observed. The data includes absences due to any illness, any respiratory illness and other more specific symptoms. While data are available for both...
prevailant and incident absences, our analysis concentrates on the incident absences, i.e. the first day of absence following a school day. A school attendance of a preceding school day is required for eligibility into the risk pool. Hence, the number of people at risk on any given day excludes those students from schools not in session and prevalent absentees. All days following a holiday were also recoded as Mondays in order to properly account for the possible mixes of prevalent and incident absences.

The daily pollution levels for $O_3$ and $PM_{10}$ were obtained from central community monitors in each of the 12 communities that were established for the larger cohort study. For each day, the 24 h average, 10 AM to 6 PM average, and the 1 h maximum of $O_3$ and $PM_{10}$ were abstracted. In addition, the 24 h average, maximum, and minimum temperatures on the same days were available. The daily series of $O_3$ and $PM_{10}$ were highly correlated within each community (with correlation coefficients ranging between 0.43 and 0.83 for Long Beach and Lake Arrowhead respectively). The overall correlation between the daily series of $PM_{10}$ and $NO_2$ was moderately positive while $O_3$ and $NO_2$ were generally uncorrelated. The long-term average of $O_3$ was weakly correlated with $PM_{10}$ ($r = 0.1$) and $NO_2$ ($r = -0.2$). The long-term averages of $PM_{10}$ and $NO_2$ were highly correlated ($r = 0.5$).

While absence data are available by subject and day, the time series analysis aggregates over subjects and concentrates on the acute effects of air pollution and other factors on the daily incident absence rates, allowing for effects to be distributed over a certain lag period. We illustrate the use of the proposed two-stage approach by examining the effects of 10 AM to 6 PM average $O_3$ and 24 h average $PM_{10}$ on school absenteeism due to any illness. These pollution metrics were chosen to be the most appropriate, on substantive grounds. A more complete analysis with a detailed discussion on the epidemiologic implications of the results is reported in Gilliland et al. (2001). The first-stage models are adjusted for long-term time trends via a smooth function of time with three degrees of freedom, daily ambient temperature and day-of-the-week effect.

In almost all models there was a very strong day-of-the-week effect, with higher incident absence rates for Monday and Tuesday. The effect of daily ambient 24 h average temperature was not found to be significant in most models, but all models were adjusted for temperature since it is considered to be an important confounder. The second-stage results for the overall aggregate acute effects of 10 AM to 6 PM daily $O_3$ and 24 h average $PM_{10}$ for illness-related absences are summarized in Table 1.

According to our AIC criterion, the second-degree polynomial with 25 day lag (AIC = 962.3) and the third-degree polynomial with 29 day lag (AIC = 962.9) were chosen to be the best. While the second-degree polynomial with 25 day lag is more parsimonious, we wanted to allow for flexibility in some communities that exhibited more curvature (than permitted by a parabola) and hence we chose the 30 day lag with third-degree polynomial. An ideal alternative would have been to search for community-specific lag and degree combinations, but our absence monitoring data has short time series with some gaps and hence was not suitable for this more ambitious scheme. We also present results from 15 day lag models for comparison. For illness related absences, the estimate of the effect of ozone on the incident absence rate, summed over the 30 day lag period and aggregated over the 11 communities, is found to be 0.024. After scaling by the minimum of the community-specific inter-quartile ranges, 20 ppb, this amounts to a 61.6% increase in the incident absence rate.

Table 2 shows the results from the first stage estimates of autocorrelation and overdispersion parameters. Note that the degree of autocorrelation varies by communities, and hence separate and explicit estimation of these parameters gives additional insight into the characteristics of the data. For example, the degree of autocorrelation in Long Beach is quite weak ($\hat{\alpha}_1 = 0.01$ and $\hat{\alpha}_2 = 0.02$ for the second-order autoregressive model with 30 day lag period), while Alpine exhibits a somewhat stronger degree of autocorrelation ($\hat{\alpha}_1 = 0.33$ and $\hat{\alpha}_2 = 0.05$ for the same model). The estimates of the constant overdispersion factor varied from 1.83 for Miraloma to 6.8 for Lake Arrowhead.

The lag-specific trend of these aggregated effects are shown pictorially in Figures 1–3. Figure 1 depicts the lag-specific ‘acute’ effects, in per cent differences per 20 ppb, averaged over the 11 communities (panel
A two-stage model for multiple time series data of counts

Table 1. Acute effects (overall and stratified by long-term levels of $O_3$ and $PM_{10}$) of 10 AM–6 PM average $O_3$ and 24 h average $PM_{10}$ on illness-related absence rates aggregated over a specified lag period. Models are based on a second-order autoregressive model

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Lag (degree)</th>
<th>Stratum</th>
<th>Estimate (s.e.)</th>
<th>$Z$-value</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O_3$</td>
<td>15 (2)</td>
<td>Overall</td>
<td>0.007 (0.006)</td>
<td>1.02</td>
<td>0.309</td>
</tr>
<tr>
<td></td>
<td>30 (3)</td>
<td>Overall</td>
<td>0.024 (0.014)</td>
<td>1.78</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>30 (3)</td>
<td>Low $O_3$, Low $PM_{10}/NO_2$</td>
<td>0.066 (0.027)</td>
<td>2.49</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High $O_3$, High $PM_{10}/NO_2$</td>
<td>$-0.009$ (0.028)</td>
<td>$-0.33$</td>
<td>0.744</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low $O_3$, Low $PM_{10}/NO_2$</td>
<td>0.029 (0.037)</td>
<td>0.78</td>
<td>0.438</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High $O_3$, High $PM_{10}/NO_2$</td>
<td>0.013 (0.017)</td>
<td>0.77</td>
<td>0.439</td>
</tr>
<tr>
<td>$PM_{10}$</td>
<td>15 (2)</td>
<td>Overall</td>
<td>0.007 (0.012)</td>
<td>0.60</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>30 (3)</td>
<td>Overall</td>
<td>0.022 (0.025)</td>
<td>0.87</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>30 (3)</td>
<td>Low $O_3$, Low $PM_{10}/NO_2$</td>
<td>0.038 (0.048)</td>
<td>0.79</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High $O_3$, High $PM_{10}/NO_2$</td>
<td>0.011 (0.048)</td>
<td>0.22</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low $O_3$, Low $PM_{10}/NO_2$</td>
<td>0.030 (0.077)</td>
<td>0.39</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High $O_3$, High $PM_{10}/NO_2$</td>
<td>0.016 (0.032)</td>
<td>0.49</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Table 2. Community-specific estimates of second-order autoregressive IWFLS model parameters in a 30 day third-degree polynomial distributed lag model for 10 AM to 6 PM average $O_3$ and illness absence rates. Community ranking is given according to low versus high with respect to the median of long-term pollution averages

<table>
<thead>
<tr>
<th>Community</th>
<th>Ranking by $(O_3, PM_{10}/NO_2)$</th>
<th>AR-2 Estimates $(\hat{\alpha}_1, \hat{\alpha}_2)$</th>
<th>Over-dispersion factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpine</td>
<td>(H, L)</td>
<td>(0.330, 0.048)</td>
<td>1.92</td>
</tr>
<tr>
<td>Lake Elsinore</td>
<td>(H, H)</td>
<td>(0.301, 0.006)</td>
<td>2.37</td>
</tr>
<tr>
<td>Lake Arrowhead</td>
<td>(H, L)</td>
<td>(0.111, 0.148)</td>
<td>6.80</td>
</tr>
<tr>
<td>Lancaster</td>
<td>(L, L)</td>
<td>(0.121, 0.138)</td>
<td>3.75</td>
</tr>
<tr>
<td>Lompoc</td>
<td>(L, L)</td>
<td>(0.116, 0.016)</td>
<td>1.99</td>
</tr>
<tr>
<td>Long Beach</td>
<td>(L, H)</td>
<td>(0.011, 0.016)</td>
<td>1.60</td>
</tr>
<tr>
<td>Miraloma</td>
<td>(L, L)</td>
<td>(0.139, 0.008)</td>
<td>1.83</td>
</tr>
<tr>
<td>Riverside</td>
<td>(H, H)</td>
<td>(0.173, 0.261)</td>
<td>2.86</td>
</tr>
<tr>
<td>San Dimas</td>
<td>(H, H)</td>
<td>$(-0.015, 0.098)$</td>
<td>4.27</td>
</tr>
<tr>
<td>Santa Maria</td>
<td>(L, L)</td>
<td>(0.160, 0.059)</td>
<td>3.46</td>
</tr>
<tr>
<td>Upland</td>
<td>(H, H)</td>
<td>(0.083, 0.010)</td>
<td>2.13</td>
</tr>
</tbody>
</table>

(b)), along with the community-specific and lag-specific effects (panel (a)). The estimates of acute effects of $O_3$ by stratifying the 11 communities into low/high categories according to their long-term $O_3$ and $PM_{10}/NO_2$ profiles are also given in Table 1. The results indicate that the acute effect of $O_3$ is most evident in the communities with low long-term averages of both $O_3$ and $PM_{10}/NO_2$. For the second-order autoregressive 30 day polynomial distributed lag model, there is a 274.3% increase in the illness related absence rates for 20 ppb increase in daily $O_3$ levels.

Figure 2 shows the trend in the ‘acute’ effects of $O_3$ for a second-degree polynomial distributed lag model over a 15 day period. The general trend is similar. The 5–15 day lag period appears to be associated
with increases in the incident absence rates. One possible explanation for this lag structure is that increases in O$_3$ levels may increase the susceptibility for illness or infection and hence it takes a few days for it to be manifested via school absenteeism.

The lower half of Table 1 shows that the ‘acute’ effect of PM$_{10}$ was not significant, even though there was a slight increase at short lag periods as depicted in the plots given in Figure 3. There was also no significant ‘acute’ effect of PM$_{10}$, when communities were stratified by long-term levels of PM$_{10}$/NO$_2$.

The estimates of the ‘chronic’ effects of pollution based on model (5) are summarized in Table 3. These results indicate that increments in long-term levels of PM$_{10}$ are significantly associated with increases in incident rates of illness-related school absenteeism. It is interesting to note that while ‘acute’ fluctuations of O$_3$ are associated with increase in illness related absenteeism (Table 1), there are no significant associations between long-term O$_3$ levels and illness-related school absenteeism.
A two-stage model for multiple time series data of counts

Fig. 2. Acute effects of $\text{O}_3$, per 20 ppb, on illness absences summarized over the 11 communities. Results are based on community-specific second-degree polynomial distributed lag models over a 15 day lag period.

Fig. 3. Acute effects of $\text{PM}_{10}$, per 10 $\mu g\, m^{-3}$, on illness absences summarized over the 11 communities. Results are based on community-specific third-degree polynomial distributed lag models over a 30 day lag period.
Table 3. ‘Chronic’ effects of 10 AM to 6 PM average $O_3$ and 24 h average $PM_{10}$ on illness-related absence rates from a 30 polynomial distributed lag model. Models are based on a second-order autoregressive model

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Estimate (s.e.)</th>
<th>Z-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O_3$</td>
<td>0.002 (0.015)</td>
<td>0.16</td>
<td>0.88</td>
</tr>
<tr>
<td>$PM_{10}$</td>
<td>0.026 (0.009)</td>
<td>2.81</td>
<td>0.02</td>
</tr>
</tbody>
</table>

To check model fit, the models were fitted by assuming autoregressive models of up to order 5. There was little change in the results beyond AR-2 and hence autoregressive models of order 2 were used in the final models. Plots of the community-specific Pearson residuals revealed that the amount of overdispersion was clearly different among the communities, and the correction for community-specific overdispersion worked well. For example, Lake Arrowhead exhibited higher variability in the Pearson residuals with some very large values, but the residual plot after correction for overdispersion behaved much better and revealed no remaining patterns. The high overdispersion factor for some of the communities such as Lake Arrowhead (Table 2) might indicate that the overall model may not be adequate for that community, but we believe that the community-specific correction for overdispersion and autocorrelation has compensated for this shortcoming.

4. DISCUSSION

The analysis of the absence monitoring data suggests that the incidence rate of illness-related absences is positively associated with any daily increments in the 10 AM to 6 PM $O_3$ levels. This effect is more pronounced in the communities that generally have low levels of $O_3$, $PM_{10}$ and $NO_2$. While there is no significant association between daily $PM_{10}$ increments and illness-related school absences, communities that have higher long-term levels of $PM_{10}$ tended to have more illness-related school absences. School absenteeism provides a tangible measure of how respiratory symptoms that are induced and/or exacerbated by air pollution affect the well being and educational development of children. Hence, it is potentially of significant public health importance in setting air quality standards.

In this paper, we have proposed a two-stage model for time series data of counts from multiple communities. The proposed method allows for the explicit estimation of the level of autocorrelation in the individual communities, levels of overdispersion in the individual communities, common estimates for confounding variables and inference on the appropriately lagged (via community-specific polynomial distributed lag modeling) acute effects of pollution aggregated over all communities or stratified by clusters of communities according to some criterion (e.g. long-term levels of the same pollutants or another ecologic factor). Note that our two-stage model is based on the iteratively weighted and filtered least-squares approach of Zeger (1988), while the papers by Dominici et al. (2000) and Schwartz (2000b) are based on a generalized additive model approach that implicitly accounts for the autocorrelation and overdispersion in the data via a smooth function of time. A systematic comparison of the two modeling approaches would be very useful and could potentially provide more insight into the modeling process. We believe that the explicit modeling of the autoregressive structure and overdispersion provides important information on the way air pollution operates in the various communities. The overdispersion estimates also show how the data from each of the communities deviates from the nominal Poisson dispersion. A strong level of overdispersion would indicate that there might be individual confounders that are not being accounted for in the models, so proper account for this unexplained variation would become important.
The selection of the number of lag days and degree of polynomial to be used in the distributed lag models needs further attention. Ideally, one would proceed in a sequential manner in the combined mixed effects model framework by starting with a model that has a relatively large number of lag days for a given degree. One would then use sequential testing procedures to select the most appropriate number of lag days over a grid of values for the degree of the polynomial.

The proposed methods are applicable to any ecologic study that examines the relationship between a daily outcome of counts (such as daily mortality, daily morbidity and school absenteeism) and daily environmental exposure (such as daily ambient air pollution, daily pollen counts, etc.) from each of multiple communities, and combine evidence across communities in a meaningful way. Note that it is also possible to directly analyze the binary time series data that arise at the individual level, probably along the lines of earlier methods described in Cox and Snell (1989). This could circumvent the need to aggregate data by day or subject. It could also potentially provide insight into the exposure–response relationship. Developments along these lines and methods that account for any between-community spatial correlation will be reported elsewhere.

5. ACKNOWLEDGEMENTS

We thank Frank Gilliland for valuable discussions on the data analysis and Dan Stram for helpful discussions on the second-stage mixed-effects models. This research was supported by the California Air Resources Board (Contract A033-186), the National Institute of Environmental Health Science (Grants 5P3OES07048-04, 1PO1ES093581-01) and the Environmental Protection Agency (Contract CR824034-01-3, Grant R826708-01-0).

REFERENCES


[Received 27 July, 2000; revised 10 November, 2000; accepted for publication 22 November, 2000]