Seasonal adjustment methods for the analysis of respiratory disease in environmental epidemiology

Bircan Erbas[1] and Rob J Hyndman[2]   9 August 2000

Abstract: We study the relationship between daily hospital admissions for respiratory disease and various pollutant and climatic variables, looking particularly at the effect of seasonal adjustment on the estimated models.

Often time series exhibit seasonal behaviour and adequate control for the presence of a seasonal component is essential before one attempts to model the complex pollution-health association. We show that if these factors are not adequately controlled for, spurious effects of pollutants and climate on morbidity/mortality can be induced.

We present a method of seasonal adjustment called STL (Seasonal-Trend decomposition based on Loess smoothing), and apply it to pollution and climate data. We will use the seasonally adjusted series in a Generalized Linear Models and Generalized Additive Models analysis of the effects of pollution and climate on hospital admissions for Chronic Obstructive Pulmonary Disease in Melbourne, Australia for the period 1989–1992.

---

[1]Department of General Practice & Public Health, The University of Melbourne, VIC 3010, Australia.
[2]Department of Econometrics and Business Statistics, Monash University, VIC 3800, Australia.

Introduction

The presence of a "long-wave length" pattern in hospital admissions and mortality for Respiratory Disease has been a common methodological issue in many studies[1,2,3]. These "long-wave length" patterns are commonly known as seasonality in time series. A seasonal pattern exists when a series is influenced by a seasonal factor (e.g., day of week, or month of year)[4].

Seasonal patterns in the response (hospital admissions and mortality for respiratory disease) have been commonly estimated using Fourier series terms[5,6,7]. Although this method of modelling seasonality works well in adequately capturing the underlying seasonal pattern in the response, it doesn't accommodate the possible seasonal pattern in each of the explanatory variables. When there is seasonality in the explanatory variables, there will inevitably be colinearity which may lead to spurious conclusions concerning the effect of an individual pollutant, and makes it difficult to separate the effects of pollutants. Despite these serious difficulties, accounting for the presence of a seasonal pattern in the pollutants and climatic data has been largely ignored.

Our approach will be to first seasonally adjust all explanatory variables. This greatly reduces the colinearity problem, without affecting the interpretability of the model. It might seem appropriate and consistent to also seasonally adjust the response variable, however this is not advisable since it is a count variable and we wish to use a Poisson model. Instead, we include Fourier series terms to control for the seasonal pattern in hospital admissions for COPD. To seasonally adjust the pollutants and climatic series we will use a time series seasonal adjustment method called the STL (Seasonal-Trend Decomposition procedure based on Loess smoothing) method.[8].

Daily counts of hospital admissions for ICD 496 (chronic obstructive airways disease) were obtained from the Public Health & Development Division of the Department of Human Services, Victoria, Australia. Air pollution data was obtained from the Environment Protection Authority (EPA), which maintains a network of 12 monitoring stations around

Melbourne. Daily maximum hourly levels of Nitrogen Dioxide (NO_2), Ozone (O_3), Sulfur Dioxide (SO_2) and the Air Particles Index (API) were obtained. Daily humidity (hu) and dry bulb temperature (db) measurements were obtained from the Commonwealth Bureau of Meteorology which has four major stations in the Melbourne metropolitan area.

We will study the effect of the pollutants and climate variables on hospital admissions for Chronic Obstructive Pulmonary Disease (COPD) in Melbourne, Australia for the period 1989–1992. We will apply the STL method to seasonally adjust all variables. We will utilize Generalized Linear Models (GLMs)[9] and Generalized Additive Models (GAMs)[10] to analyze the effects of pollutants and climate on hospital admissions for COPD. We will then compare these results with those we have obtained previously[11].

Methodology

Time series seasonal adjustment

A crucial preliminary step before modeling the short term effects of pollution and climate on daily hospital admissions/mortality, is the examination of the underlying behaviour of each of these potential covariates. Since pollutants and climatic variables are time series, decomposition methods may be used to break up a time series into the following components: trend, seasonal and irregular.

The trend component consists of the underlying long-term aperiodic rises and/or falls in the level of the series over time. The seasonal component is a pattern that is recurrent over time. The irregular component is the remaining pattern in the series not attributed to trend or seasonality[4]. Both trend and seasonality are potential confounding variables in any analysis, so their identification and removal are important.

Seasonality is an established strong confounding variable in the analysis of daily hospital admissions/mortality data. Time series decomposition methods allow us to identify

the strength of the seasonal component in each of the pollutants and climatic variables. After identification, the seasonal component of the series will be removed and the resultant seasonally adjusted series will be used in subsequent analysis. Thus extracting the seasonal component will allow a clearer picture of the other features of the time series.

A number of time series decomposition methods are available. A relatively simple decomposition method is classical decomposition[4], but that has several drawbacks including bias problems near the ends of the series and an inability to allow a smoothly varying seasonal component. To overcome these difficulties, we prefer the STL (Seasonal-Trend decomposition procedure based on Loess) method. We will assume an additive decomposition:

$$Y_t = T_t + S_t + E_t$$

where Y_t denotes the time series of interest, T_t denotes the trend component, S_t denotes the seasonal component and E_t denotes the remainder (or irregular) component. The seasonally adjusted series, Y_t^* is computed simply by subtracting the estimated seasonal component from the original series, $Y_t^* = Y_t - \hat{S}_t$.

STL consists of a sequence of applications of the Loess smoother[12] to give a decomposition that is highly resistant to extreme observations[4]. The STL method involves an iterative algorithm to progressively refine and improve estimates of the trend and seasonal components. STL consists of two recursive procedures, one nested within the other, called the inner loop and the outer loop. In each iteration of the inner loop, the seasonal and trend-cycle components are updated once. An iteration of the outer loop consists of one or two iterations of the inner loop followed by an identification of extreme values. Future iterations of the inner loop downweight the extreme values that were identified in the previous iteration of the outer loop. Between 10 and 20 iterations of the outer loop are usually carried out in total.

We describe the steps for a single iteration of the inner loop assuming a series Y_t which consists of daily observations (so the seasonal period is 365). The iteration consists of updating the estimate of the trend component and calculating a new estimate of the sea-

sonal component. The whole procedure must start with some initial estimate of the trend. This is set to be zero. That is, the procedure begins by assuming no trend at all. This poor estimate is quickly updated to something more reasonable after one iteration.

Step 1 Subtract an estimated trend from the original data to obtain the detrended values $Y_t' = Y_t - \hat{T}_t$.

Step 2 For each day of the year, the detrended values are collected to construct a daily sub-series. Each of the 365 sub-series are separately smoothed by a Loess smoother. A preliminary seasonal component is constructed by connecting the smoothed sub-series back together. An estimate of the seasonal component a few days before and after the observed data are used to extrapolate the Loess smoother.

Step 3 A $3 \times 365 \times 365$ moving average is applied to the preliminary seasonal component estimated in Step 2. The result is in turn smoothed by a Loess smoother of length 365.

Applying a weighted-moving average resulted in a loss of values at both the beginning and the end of the series. However, this was overcome by the extrapolation of the seasonal component in Step 2.

The purpose of this step is to identify any trend that may have contaminated the preliminary seasonal component in Step 2. If there is little trend in the preliminary seasonal component, the result of this step will be a series with all values close to zero.

Step 4 We estimate the new seasonal component as the difference between the preliminary seasonal component in Step 2 and the smoothed seasonal component in step 3.

Step 5 The result from Step 4 is subtracted from the original series to give the seasonally adjusted series $Y_t^* = Y_t - \hat{S}_t$.

Step 6 To obtain a new estimate of the trend component T_t, we smooth (by Loess) the now seasonally adjusted series Y_t^*.

The outer loop begins with one or two iterations of the inner loop. The resulting estimates of trend and seasonal components are then used to calculate the irregular component:

$$\hat{E}_t = Y_t - \hat{T}_t - \hat{S}_t.$$

Large values of $\hat{E}_t$ indicate an extreme observation. These are identified and a weight calculated. That concludes the outer loop.

To down weight the effects of the extreme observations, future iterations of the inner loop use these weights from Step 2 to Step 6. Also, future iterations of the inner loop begin with the trend component from the previous iteration rather than starting with zero as in the very first iteration of the inner loop.

There are two Loess smoothing parameters that must be selected when using the STL procedure: the seasonal smoothing parameter used in Steps 2 and Steps 6, and a smoothing parameter for the trend component calculated in Step 6 of the inner loop. Both smoothing parameters determine the variation from year to year in the seasonality and the trend. Small smoothing parameters allow substantial variation from year to year and a large smoothing parameter allows very little variation from year to year.

The procedure is implemented using S-Plus 2000 for Windows.

Generalized linear and additive models

Generalized Linear Models (GLMs) with quasi-likelihood estimation[9] to model the overdispersion frequently encountered in hospital admissions data[3] were applied to COPD hospital admissions in Melbourne, Australia for the period 1989–1992. Potential explanatory variables are the seasonally adjusted nitrogen dioxide, ozone, humidity, dry bulb temperature and air particle index, and the non seasonally adjusted sulfur dioxide. We also included day of week dummy variables, and Fourier series terms (i.e., $\cos(2\pi kt/365)$ and $\sin(2\pi kt/365)$ for $k = 1, 2, 3, 4$) to control for the seasonal effect in COPD hospital admissions. Lags of up to 2 days were included in the analysis for each

pollutant and climatic variable.

Covariates were selected using an efficient step-wise selection process in S-Plus using Akaike's Information Criterion[13] (AIC) to evaluate different models. The model with the smallest AIC was chosen as the final model.

A non parametric alternative to the GLM is the Generalized Additive Model[10] (GAM). These models allow each of the explanatory variables to enter the model in a non-linear manner. As for the GLM, we used step-wise selection in S-Plus, to choose the covariates, selecting the model with the smallest AIC statistic. The same explanatory variables were used as for the GLM, except that we allowed each of them to enter the model non-linearly. The non-linear functions were estimated using cubic smoothing splines with four degrees of freedom. The AIC was used to determine whether a variable should be included in the model using a spline or as a linear function.

## Results

Seasonal adjustment

Figure 1 displays a time series plot of the response series (COPD) and each of the pollutants and climatic series. There is evidence of seasonality in all but SO_2. For each series, we determined the length of the seasonal and trend window for the loess smoothing by trial and error, selecting values which appear reasonable for the data. We also assume a "periodic" component for seasonality; that is, there is the same cycle for each year of the series.

We use decomposition plots[4,14] to help visualize the decomposition procedure. A decomposition plot displays a time series plot of the original data in the top panel and the remaining panels provide a plot of the trend, seasonal pattern and remaining variation which is not accounted for by the trend and seasonality.
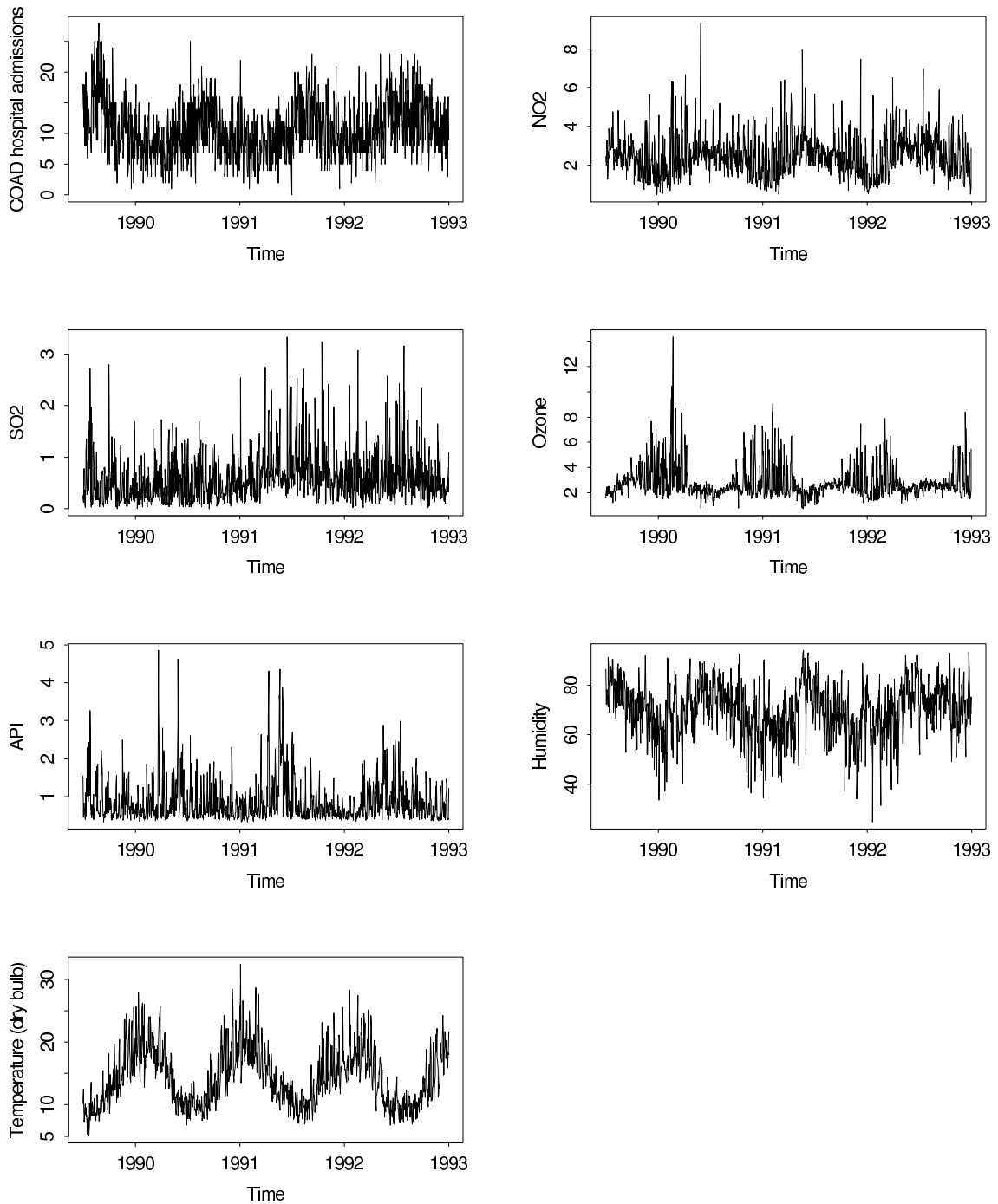
Figure 1: *A time plot of each series in the data set, for the time period 1 July 1989 to 31 December 1992*
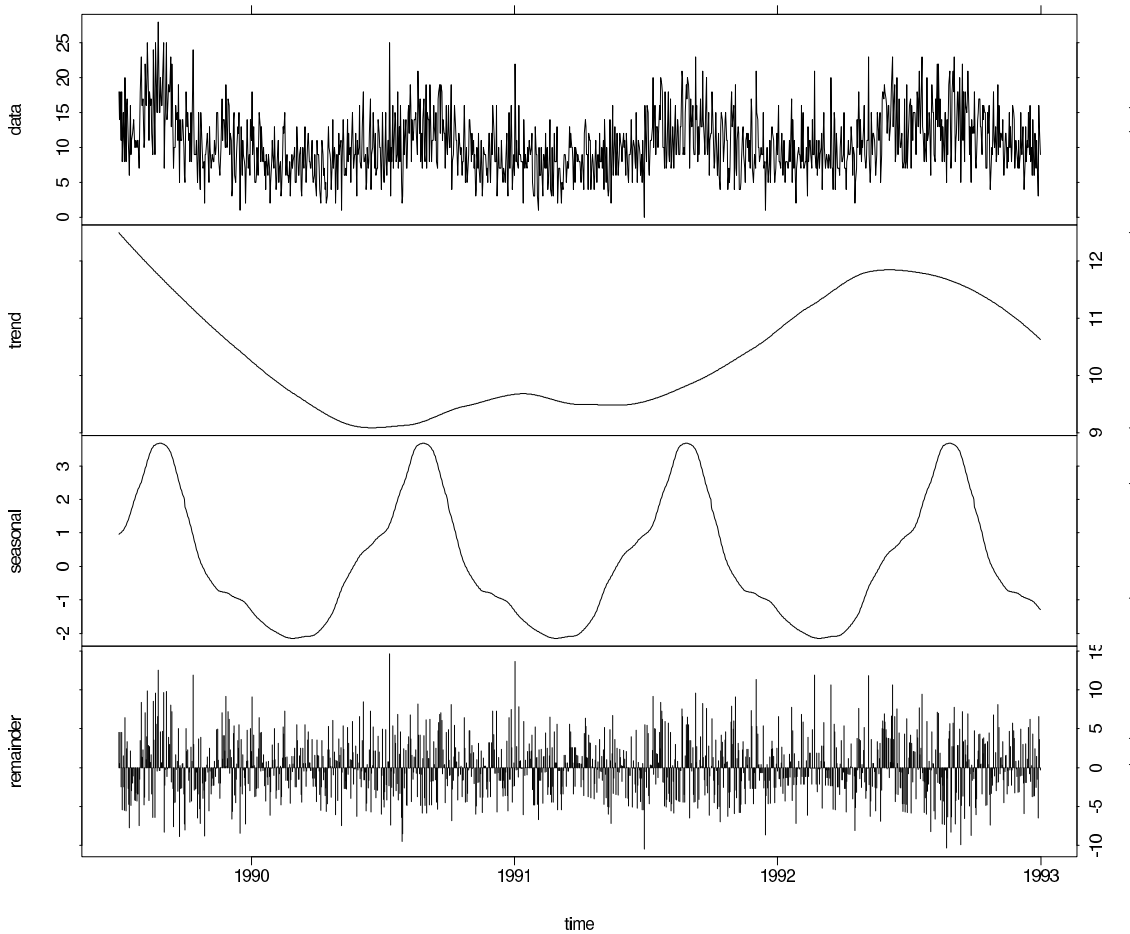
Figure 2: *A decomposition plot of COPD, for the time period 1 July 1989 to 31 December 1992*

In Figure 2 we display the decomposition plot for COPD as an example. The length of the bars on each side of the decomposition plots are an indication of the strength of the individual components. Each bar is the same length, but plotted on different scales. Clearly there is very little trend in the series (indicated by the long bar in the trend plot) but substantial seasonality present (indicated by the shorter bar in the seasonal plot).

A scatter plot matrix is an exploratory graphical method introduced by Chambers et al.[15] (1983) to investigate more than two series in an multi-dimensional space. In Erbas & Hyndman[11] (2000) we used a scatter plot matrix to show that several supposed non-linear relationships between hospital admissions for COPD and the pollutant and climatic covariates were, in fact, induced by seasonality.
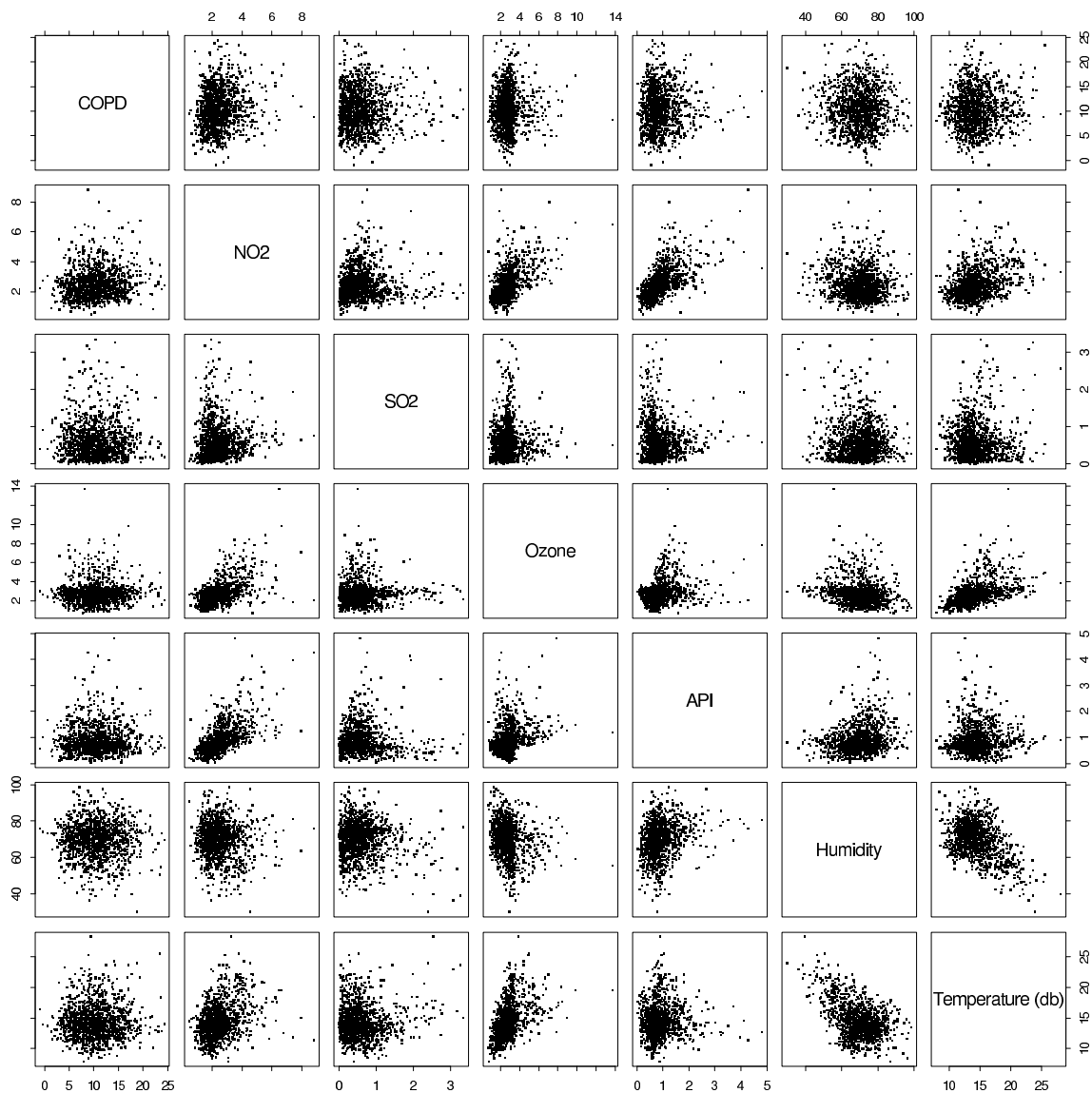
Figure 3: *Pairwise scatter plots for hospital admissions for COPD, pollutants and climate. All variables seasonally adjusted except SO$_2$.*

Figure 3 displays a scatter plot matrix of the data after all covariates (including COPD) have been seasonally adjusted. It is difficult to visualize the non-linear relationships between COPD hospital admissions and climate reported in previous studies[16,17,18]. However, we shall see that there is some non-linearity between COPD and db temperature, and COPD and SO$_2$, both covariates lagged by two days.

Table 1: *Regression coefficients, corresponding standard errors and p values obtained by a GLM analysis*

Parameter	Estimate	Standard Error	p-value
Intercept	2.4397	0.0821	0.0000
$NO_{2,t}$	0.0232	0.0103	0.0245
hu_t	-0.0026	0.0011	0.0151
t	1.0821	0.3331	0.0012
t^2	2.4963	0.3601	0.0000
D_1	-0.0420	0.0172	0.0145
D_2	-0.0163	0.0101	0.1058
D_3	-0.0057	0.0071	0.4250
D_4	-0.0064	0.0056	0.2525
D_5	-0.0505	0.0052	0.0000
D_6	-0.0216	0.0042	0.0000
$\sin(2\pi t/365)$	0.1719	0.0150	0.0000
$\cos(2\pi t/365)$	0.1515	0.0141	0.0000
$\sin(4\pi kt/365)$	0.0461	0.0143	0.0013
$\cos(4\pi kt/365)$	-0.0011	0.0143	0.9385
$\sin(6\pi kt/365)$	-0.0208	0.0143	0.1468
$\cos(6\pi kt/365)$	-0.0456	0.0142	0.0014
$\sin(8\pi kt/365)$	-0.0390	0.0140	0.0054
$\cos(8\pi kt/365)$	-0.0170	0.0139	0.2208

Generalized linear models

Table 1 displays the estimation results of a GLM analysis with hospital admissions for COPD as the response and day of week dummies, a quadratic time trend, and all other explanatory variables were included linearly in the model.

A GLM analysis for COPD hospital admissions in Melbourne, Australia for the period 1989–1992 was previously reported in Erbas & Hyndman[11] (2000). However, the GLMs didn't include a seasonal adjustment of the explanatory variables that exhibited an underlying seasonal pattern. In our previous analysis, we reported a similar GLM except that our previous model included API lagged at 1 and 2 days. All other variables are the same. Thus, the effect of API appears to be spurious and induced by seasonality.

Table 2: *Regression coefficients of linear terms, corresponding standard errors and p values obtained by a GAM analysis*

Parameter	Estimate	Standard Error	p-value
Intercept	2.1362	0.1365	0.0000
$NO_{2,t}$	0.0228	0.0102	0.0252
$ozone_{t-2}$	0.0217	0.0104	0.0371
api_{t-2}	-0.0400	0.0205	0.0510
hu_t	-0.0027	0.0011	0.0129
hu_{t-2}	0.0029	0.0013	0.0202
D_1	-0.0407	0.0170	0.0167
D_2	-0.0146	0.0100	0.1436
D_3	-0.0047	0.0071	0.5050
D_4	-0.0068	0.0055	0.2207
D_5	-0.0507	0.0051	0.0000
D_6	-0.0226	0.0041	0.0000
$\sin(2\pi t/365)$	0.1737	0.0144	0.0000
$\cos(2\pi t/365)$	0.1596	0.0151	0.0000
$\sin(4\pi kt/365)$	0.0331	0.0143	0.0206
$\cos(4\pi kt/365)$	-0.0015	0.0141	0.9138
$\sin(6\pi kt/365)$	-0.0234	0.0141	0.0980
$\cos(6\pi kt/365)$	-0.0495	0.0140	0.0004
$\sin(8\pi kt/365)$	-0.0417	0.0140	0.0029
$\cos(8\pi kt/365)$	-0.0167	0.0137	0.2240

Generalized additive models

We obtained the following GAM for the data:

$$
\begin{aligned}
\mathrm{E}(Y_t|X_t) \;=\; & \exp\Big\{\beta_0 + \beta_1 NO_{2,t} + \beta_2 ozone_{t-2} + g_3(SO_{2,t-2}) + \beta_4 api_{t-2} \\
& + g_5(db_{t-2}) + \beta_6 hu_t + \beta_7 hu_{t-2} + g_8(t) \\
& + \beta_9 D_1 + \beta_{10} D_2 + \beta_{11} D_3 + \beta_{12} D_4 + \beta_{13} D_5 + \beta_{14} D_6 \\
& + \sum_{k=1}^{4} \left[\gamma_k \cos(2\pi kt/365) + \theta_k \sin(2\pi kt/365) \right] \Big\}
\end{aligned}
\tag{1}
$$

where $Y_t|X_t$ is Pseudo-Poisson (i.e., Poisson with overdispersion). Note that SO_2 lagged 2 days, db_t lagged 2 days, and time were modelled using a non-linear (spline) function. All other variables are included linearly.

Table 2 displays the coefficients for the linear terms in model 1.

This differs from the model reported in Erbas & Hyndman[11] (2000) in that humidity and dry bulb temperature (lagged 2 days) are now significant and included. All other variables are the same. Thus, it appears that the seasonality in humidity and dry bulb temperature was masking their importance in our previous analysis, and that the seasonal adjustment done here has led to their inclusion.

Apart from the non-linear time trend, two covariates were selected to be non-linear: dry bulb temperature and sulphur dioxide (both lagged by 2 days). Figure 4 depicts these relationships.
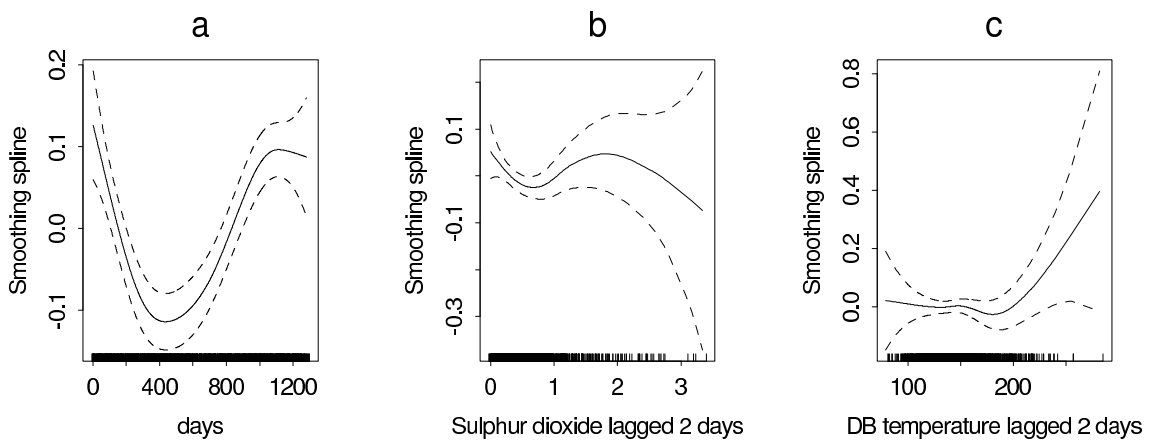
Figure 4: *Non-linear functions in the generalized additive model (1), fitted using cubic smoothing splines. (a) $g_8(t)$, the smooth underlying time trend; (b) $g_3(SO_{2,t-2})$, the non-linear function of sulphur dioxide (lagged 2 days); (c) $g_5(db_{t-2})$, the non-linear function of dry bulb temperature (lagged 2 days). Dashed lines represent pointwise 95% confidence intervals.*

In the analysis of daily morbidity/mortality, serial correlation is also an important methodological issue[2,3]. Autocorrelation plots of the residuals allow a visual examination of any remaining correlation structure. We will use randomized quantile residuals[19] developed for non normal data.

An autocorrelation plot of the randomized quantile residuals from the GAM analysis in 1 is displayed in Figure 5. There is clearly very little remaining significant correlation in the residuals after both seasonally adjusting the explanatory variables and applying GAM methodology.
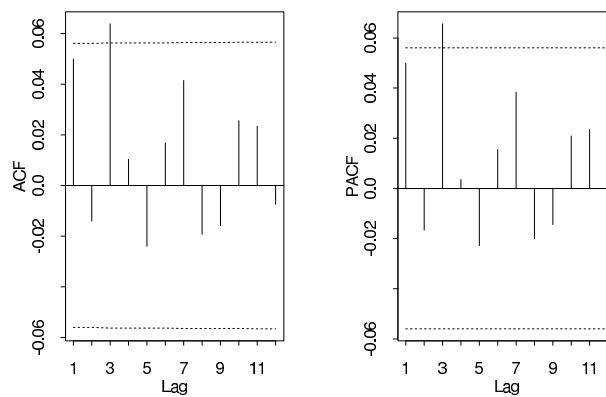
Figure 5: *Autocorrelation plot of random quantile residuals from the fit of the GAM in equation (1). The equivalent plot for the GLM showed more significant autocorrelation at lags 1, 3 and 7.*

The GLM randomized quantile residuals exhibited greater (although still small) autocorrelation which was significant at lags 1, 3 and 7. Allowing non-linear functions as in the GAM can greatly reduce the autocorrelation[11] inherent in morbidity/mortality data and thereby greatly simplify the analysis.

Discussion

The analysis described here demonstrates that the effects of pollution and climate on daily counts of morbidity/mortality can be masked by seasonality. This arises because of the colinearity between the variables which is induced by seasonality.

We argue that it is only possible to assess the effects of individual variables if this confounding is reduced via seasonal adjustment.

We have presented a time series decomposition method, namely STL, that handles any length of seasonality and can handle time series with missing values, something other seasonal adjustment methods cannot handle easily. We have used decomposition plots to break down the variables into a seasonal and a trend component. Then the explanatory variables were seasonally adjusted accordingly.

We then use a GLM and a GAM analysis with the seasonally adjusted variables to study

the effects of each variable on COPD hospital admissions. In both the GLM and GAM analysis, a linear effect of nitrogen dioxide is statistically significant. We identified a significant nonparametric smooth effect of dry bulbs temperature lagged 2 days in the GAM analysis. This was not identified in our previous analysis[11] where seasonal adjustment was not employed. We have shown that seasonality masked the true effects of temperature and humidity, since they became significant only after seasonal adjustment.

The issue of autocorrelation in the residuals was overcome in the GAM analysis, since we were able to have a combination of linear and smooth effects of the regressors. We find that allowing non-linear models can substantially reduce the autocorrelation problem in mortality/morbidity data.

We can view seasonal adjustment as a prelude to more sophisticated analysis, enabling a clearer understanding of the nature of the pollution-climate mixture, and allowing an examination of the unobscured relationships between the covariates and daily counts of morbidity/mortality.

References

1 GOLDSTEIN, I.F. & CURRIE, B. (1984) Seasonal patterns of asthma: a clue to etiology, *Environmental Research*, **33**, 201–215.

2 THURSTON, G.D. & KINNEY, P.L. (1995) Air pollution epidemiology: considerations in time series modeling, *Inhalation Toxicology*, **7**, 71–83.

3 SCHWARTZ, J., SPIX, C., TOULOUMI, G., BACAROVA, L., BARUMAMDZADEH, T., TERTRE, A LE., PIEKARSKI, T., PONCE DE LEON, A., PONKA, A., ROSSI, G., SAEZ, M., & SCHOUTEN, J.P. (1996) Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions, *Journal of Epidemiology & Community Health*, **50**(suppl 1), s3–s11.

4 MAKRIDAKIS, S., WHEELWRIGHT, S.C., & HYNDMAN, R.J. (1998) *Forecasting:*

methods & applications, 3rd ed., New York: Wiley & Sons.

5 THURSTON, G.D., ITO, K., KINNEY, P.L. & LIPPMANN, M. (1992) A multi-year study of air pollution and respiratory hospital admissions in three New York state metropolitan areas: results for 1988 and 1989 summers, *Journal of Exposure Analysis and Environmental Epidemiology*, **2**, 429–450.

6 SUNYER, J., CASTELLSAGUE, J., SAEZ, M., TOBIAS, A. & ANTO, J. (1996) Air pollution & mortality in Barcelona, *Journal of Epidemiology & Community Health*, **50**(suppl), S76–S80.

7 SIMPSON, R., WILLIAMS, G., PTEROESCHEVSKY, A., MORGAN, G., & RUTHER-FORD, S. (1997) Associations between outdoor air pollution and daily mortality in Brisbane, Australia, *Archives of Environmental Health*, **52**, 442–454.

8 CLEVELAND, R.B., CLEVELAND, W.S., MCRAE, J.E., & TERPENNING, I. (1990) STL: a seasonal-trend decomposition procedure based on Loess (with discussion), *Journal of Official Statistics*, **6**, 3–73.

9 MCCULLAGH, P. & NELDER J.A. (1989) *Generalized linear models*, London: Chapman and Hall.

10 HASTIE, T. & TIBSHIRANI, R.J. (1990) *Generalized additive models*, London: Chapman and Hall.

11 ERBAS, B. & HYNDMAN, R.J. (2000) The effect of air pollution & climate on hospital admissions for chronic obstructive airways disease: a non-parametric alternative, *Submitted*, .

12 CLEVELAND, W.S. & DEVLIN, S. (1988) Locally weighted regression: an approach to regression analysis by local fitting, *Journal of the American Statistical Association*, **74**, 596–610.

13 AKAIKE, H. (1973) Information theory & an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory*, B.N. Petrov & F.Csaki(eds), Adademiai Kidao, Budapest, 267–281.

14 ERBAS, B. & HYNDMAN, R.J. (2000) Data visualization for time series in environmental epidemiology, *Submitted*, .

15 CHAMBERS, J.M., CLEVELAND, W.S., KLEINER, B., & TUKEY, P.A. (1983) *Graphical methods for data analysis*, New York: Chapman & Hall.

16 KINNEY, P.L. & OZKAYNAK, H. (1991) Associations of daily mortality & air pollution in Los Angeles County, *Environmental Research*, **54**, 99–120.

17 SCHWARTZ, J (1995) Short term fluctuations in air pollution and hospital admissions of the elderly for respiratory disease, *Thorax*, **50**, 531–538.

18 MORGAN, G., CORBETT, S., & WLODARCYZK, J. (1998) Air pollution and hospital admissions in Sydney, Australia, 1990-1994, *American Journal of Public Health*, **88**, 1761–1766.

19 DUNN, P & SMYTH, G (1996) Randomized quantile residuals, *Journal of Computational and Graphical Statistics*, 5, 236–244.

20 BATES, D.V., BAKER-ANDERSON, M. & SITZO, R. (1990) Asthma attack periodicity: a study of hospital emergency visits in Vancouver, *Environmental Research*, 51, 51–70.

21 SAEZ, M., SUNYER, J., CASTELLSAGUE, J., MURILLO, C. & ANTO, J.M. (1995) Relationship between weather temperature & mortality: a time series analysis approach in Barcelona, *International Journal Of Epidemiology*, 24, 576-581.

22 BALLESTER, F., CORELLA, D., PEREZ-HOYOS, S., SAEZ, M. & HERVAS, A. (1997) Mortality as a function of temperature. a study in Valencia, Spain, 1991–1993, *International Journal of Epidemiology*, 26, 551–561.