1

Longitudinal Analysis of Short term Bronchiolitis Air Pollution Association using Semi Parametric Models

Sylvie Willems*, Claire Segala*, Manuel Maidenberg ** and Mounir Mesbah§

* SEPIA-SANTE, Melrand-France, ** RESPIRER, Paris-France and § LSTA, University Pierre et Marie Curie, Paris-France

Abstract:

The first aim of this work is to clearly present a very popular semi parametric methodology often used to estimate association between death or hospital counting data and pollution data, to estimate short term effects of ambient air pollution on infant bronchiolitis hospital consultations. Infant bronchiolitis is a frequent infectious disease, caused by a virus, the syncytial respiratory virus (RSV). Normally, the contact with this virus is responsible for a cold, but in infant and in some circumstances, especially at the beginning of winter, the virus can be responsible for a severe respiratory disease which lead to numerous hospital consultations and hospitalizations. A critical comparison of its practical application using S-Plus, R or SAS Proc Gam is performed. It appear that more work is needed to get a satisfactory implementation of the Schwartz method in SAS with similar results as those in S-Plus or R.

Keywords and phrases: GAM model, Air pollution, semi parametric models.

1.1 Introduction

Time-series studies of air pollution and health estimating associations between day-to-day variations in air pollution concentrations and day-to-day variations in adverse health outcomes have been widely used since the 1980s and have motivated reassessment of air quality standard in United States and Europe.

In the last 10 years, many advances have been made in the statistical modelling of air pollution time-series studies. Standard regression methods used initially have been replaced by semi-parametric approaches. Use of generalized additive models (GAMs, Hastie and Tibshirani (1990)) became popular in the mid-1990s. In a recent issue of American Journal of Epidemiology, Dominici et al. (2002) discuss the fact that the gam default convergence criteria defined in S-Plus version 3.4, (and, to a lesser degree in SAS) were not sufficiently rigorous for these analysis; the result was an overestimation of the effect of air pollution on health. More recently, in a issue of Epidemiology, Ramsay et al., (2003) point out that S-Plus and SAS functions gam use a computational approximation which, in the presence of a large correlation between the nonlinear functions included in the model (called concurrity), can underestimate the standard errors of the relative rates. The community of air-pollution researchers is now faced with the obligation of repeating analysis that have used the gam function in S-Plus and
considering further methodological issues (Health Effect Institute, 2003)). Researchers over the past decade have found other ways to fit GAM. Penalized regression splines (P-Splines) using R software are an example of technique with similar characteristics to smoothing splines, that require much less computation for standard errors. From 1997 to 2001, daily mean of environmental variables including: pollution data and meteorological data were gathered. In order to evaluate the impact of the gam problems, data are analyzed using different methods:
- GAM using LOESS functions and the default convergence parameters (using S-Plus and SAS softwares)
- GAM using more stringent convergence parameters than the default setting (in S-Plus) and GAM using P-Splines (in R).

The aim of this work is to estimate short term effect of PM10 on infant bronchiolitis hospital consultations. Infant bronchiolitis is a frequent infectious disease, caused by a virus, the syncethial respiratory virus (RSV). Normally the contact with this virus is responsible for a cold, but in infants and in some circumstances, especially at the beginning of winter, the virus can be responsible for a severe respiratory disease which lead to numerous hospital and consultations and hospitalizations. The bronchiolitis and environmental data are first presented in section 1.2. Secondly, the methods used are detailed. GAMs are described in section 1.3.1. The particular strategy applied in case of air pollution time-series studies is then exposed (section 1.3.2). Next, in section 1.3.3, criticism about the use of standard statistical software to fit GAM is displayed. Results of the bronchiolitis study are presented in section 1.4.

1.2 Material

The study consists in a longitudinal data analysis based on ecologic sanitary and environmental data.

1.2.1 Sanitary data

For 43 hospitals of the Paris region, the number of hospital consultations have been provided by the ERBUS ("Épidémiologie et Recueil des Bronchiolites en Urgence pour Surveillance") network for the period between 1997 and 2000, (Thélot et al., 1998), given a standardized definition. Bronchiolitis is defined as respiratory dyspnea and/or sibilants and wheezing for an infant of less 3 years old during the supervision period. Particularly, the counts are available for the period between the 15th of October and the 15th of January for each year. In 1999-2000, some hospitals stopped momentarily the data collection. To avoid missing values problems, only the 34 hospitals which provided complete data are retained for the study.

1.2.2 Environmental data

From 1997 to 2001, daily means of PM10 (particles with an aerometric diameter less than 10 microg) and meteorological data were gathered. Air pollutants are routinely measured at the stations of the AIRPARIF network. We retained PM10 data from the nine urban background monitoring sites which are representative of ambient air pollution in the geographical area (greater Paris). The daily average (SE) during the study period was 24.2 (10.4) micro/m3. Weather data corresponding to Paris and its outer suburbs were provided by Meteo France. Twelve covariates were transmitted for the period beginning the 01/01/1996 and finishing the
31/12/2000. Five of them were chosen using the results of a principal component analysis previously performed (Tual, 2003). The idea was to retain the covariates which bring the most information about the principal components. The weather factors retained for this study are the daily minimal temperature, the relative humidity, the daily precipitations, the daily average wind strength and the pressure at the level of the ocean.

1.3 Methods

1.3.1 Generalized additive models

GAMs (Hastie and Tibshirani (1990), Xiang, (2002)) assume that the mean of the dependent variable depends on an additive predictor through a nonlinear function. Let $Y$ be a response random variable and $X = (X_1, ..., X_p)$ be a set of predictor variables. The standard linear regression model assumes the expected value of $Y$ has a linear form and can be written as

$$E(Y|X) = f(X_1, ..., X_p) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$  \hspace{1cm} (1.1)

The additive model generalizes the linear model by modelling the expected value of $Y$ as

$$E(Y|X) = f(X_1, ..., X_p) = s_0 + s_1(X_1) + ... + s_p(X_p)$$  \hspace{1cm} (1.2)

where $s_i(X), i = 1, ..., p$ are smooth functions. These functions are not given a parametric form but instead are estimated in a nonparametric fashion. GAMs extend traditional linear models in another way, namely by allowing for a link between $f(X_1, ..., X_p)$ and the expected value of $Y$. Hence, GAMs consist of a random component, an additive component and a link function relating these two components. The response $Y$, the random component, is assumed to have a density in the exponential family:

$$f_Y(y; \theta; \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y; \phi) \right\},$$  \hspace{1cm} (1.3)

where $\theta$ is called the natural parameter and $\phi$ the scale parameter. The additive component is the quantity $\eta$ defined as

$$\eta = s_0 + s_1(X_1) + ... + s_p(X_p),$$  \hspace{1cm} (1.4)

where $s_i(X), i = 1, ..., p$ are smooth functions. Finally, the relationship between the mean $E(Y|X)$ of the response variable $Y$ and $\eta$ is defined by a link function $g(E(Y|X)) = \eta$.

Scatterplot smoothing functions, commonly referred to as smoothers, are central to GAMs. A smoother is a mathematical technique for approximating an observed variable $Y$ by a smooth function of one (or several) independent variables(s). Some examples of commonly used smoothers are smoothing splines, regression splines, LOESS functions (locally estimated polynomial regression) and kernel smoothers. These methods are called nonparametric because they make no parametric assumption about the shape of the function being estimated. Each smoother has a parameter that determines how smooth the resulting function will be. For LOESS function, that parameter is called span. For smoothing splines and natural splines, the degree of smoothing can be specified through the degrees of freedom parameter. In general, the amount of smoothing selected will have more impact than the type of smoother chosen. Further information about smooth functions can be found in Hastie and Tibshirani(1990).
GAMs have the advantage that they allow greater flexibility than traditional parametric modelling tools. They relax the usual parametric assumption and enable to uncover hidden structure in the relationship between the independent variables and the dependent variable. The link amounts to allowing for an alternative distribution for the underlying random variation besides just the normal distribution. GAMs can therefore be applied to a much wider range of data analysis problems. In particular, they are widely used for air pollution time-series studies.

1.3.2 Air pollution time-series studies strategy

A time-series analysis of air pollution and health raises distributional and modelling issues. On any given day, only a small portion of the population consults or is hospitalized. This number is a count, which suggests that a Poisson process is the underlying mechanism being modelled.

Analysis of the health effects of air pollution must account for other time-varying factors that may affect health outcomes to avoid taking the effects of such factors for pollution effects. Indeed, the basic issue in modelling is to control properly for potential confounding. Many variables show systematic variation in time. Since any two variables that show a long term trend must be correlated, searches for correlations that are more likely to be causal must exclude these trends. A second common attribute of many variables that evolve over time is seasonality. These variations would be present even if these factors were not causally related, and will induce correlations among them. Again, to focus on possibly causal associations, it is necessary to remove these patterns. A final systematic component that may bias time series regressions involves calendar specific days as day of week or holiday effects.

After season and trend, weather terms are the most important covariates to enter the model. The variables will be there to carry information about the effects of short term variations in weather on health effect. The effects of all the explanatory variables may be immediate, or may occur with some lag.

Repeated measurements are likely to be dependent. In the case where two observations closer together in time are more alike than two randomly chosen observations, this is referred to as serial correlation. If the serial correlation in the outcome is due to omitted covariates or imperfectly controlled for covariates, serial correlation will be observed in the residuals of the model. Autoregressive models represent efficient schemes to take the serial correlation into account.

Consequently, there is essential to take into account time effects and serial correlation to identify and estimate the short term relation between pollution and health events without bias. The model building is done, in accordance with the methodology developed by Schwartz (1993), including step by step: long term variations (trend), medium term variations (seasonality), short term variations (calendar specific days, weather factors), short term relations with the different pollutants and autoregressive terms if necessary.

Note that, because of the high correlations between the pollutants, multipollutant model are seldom considered.

During the model building process, diagnostics plots are used to evaluate the success of the approach. First, a plot of the residuals versus time can often identify long wavelength patterns that remain. These patterns should disappear as one goes along. Secondly, a plot of the predicted outcome over time can also be quite useful. The comparison of the graph of the predicted series and the graph of the initial series allows to judge the quality of the model. Finally, a graph of the partial autocorrelation of the residuals of the model is very important.
The sum of the autocorrelations should be as near as possible from 0 and, ideally, at the end of the analysis, the autocorrelation function of the residuals should be a white noise. In case of morbidity, the autocorrelation is also due to intrinsic factors and is more difficult to suppress. Sometimes, an important residual autocorrelation remains on the first lags. It is however necessary to obtain a white noise beyond the ten first lags as well as a reduction of the autocorrelation on the first lags.

Let $Y$, the response variable studied, $date$, the covariate representing the time (in days) and $X_1, \ldots, X_k$, the different weather factors. Note $pol$, the pollutant considered, $J_1 = I(Sunday)$, $J_6 = I(Friday)$, $\ldots$, $F = I(official\; holiday)$, $V = I(holidays)$, where $I$ is the indicator function. Let $X = (date, X_1, \ldots, X_k, pol, J_1, \ldots, J_6, F, V)$ the set of all the covariates. Note $lo_{sp_{date}}(date)$, the LOESS function for $date$ with a span equal to $sp_{date}$, and for $1 \leq i \leq k$, $lo_{sp_i}(X_i)$, the loess function for the weather factor $i$ and a span equal to $sp_i$. Let $lo_{sp_{pol}}(pol)$, the loess function corresponding to the pollutant (with a span equal to $sp_{pol}$) and verifying

$$lo_{sp_{pol}}(pol) = \beta_{pol} + f(pol), \quad (1.5)$$

with $f(pol)$, a non linear function of the pollutant, and $\beta_0$, $\beta_{11}, \ldots, \beta_{16}$, $\beta_2$, $\beta_3$ and $\beta$, the parameters of the model. The model can be written as

$$\log\{E(Y|X)\} = \beta_0 + lo_{sp_{date}}(date) + \beta_{11}J_1 + \ldots + \beta_{16}J_6 + \beta_2F + \beta_3V + lo_{sp_{pol}}(pol) + \sum_{1 \leq i \leq k} lo_{sp_i}(X_i) = a. \quad (1.6)$$

A model differing from the preceding by the fact that the effect of the pollutant is linear is also interesting. This model has the following form:

$$\log\{E(Y|X)\} = \beta_0 + lo_{sp_{date}}(date) + \beta_{11}J_1 + \ldots + \beta_{16}J_6 + \beta_2F + \beta_3V + \beta_{pol} + \sum_{1 \leq i \leq k} lo_{sp_i}(X_i) = b. \quad (1.7)$$

Since equation (1.5), models (1.6) and (1.7) are nested and since the distribution used is the Poisson distribution, a $\chi^2$-test can be performed to compare them. Indeed, the deviance corresponding to a Poisson distribution is given by $2 \sum_i \left\{ y_i \left( \log\left( \frac{\mu_i}{y_i} \right) - (y_i - \mu_i) \right) \right\}$. The difference of the deviances of models (1.6) and (1.7) has, therefore, the following form:

$$S = 2 \sum_i \left\{ y_i f(pol) - \left( \exp(a) - \exp(b) \right) \right\}. \quad (1.8)$$

The statistic $S$ follows a Chi-square distribution. If the two models are non significantly different, model (1.7) will be preferred for interpretation easiness.

At this step, some autocorrelation terms can be introduced in the model if it is necessary. Let $AR_1, \ldots, AR_j$, these terms. The final model becomes then

$$\log\{E(Y|X)\} = \beta_0 + lo_{sp_{date}}(date) + \beta_{11}J_1 + \ldots + \beta_{16}J_6 + \beta_2F + \beta_3V + \beta_{pol} + \sum_{1 \leq i \leq k} lo_{sp_i}(X_i) + \sum_{1 \leq l \leq j} AR_l. \quad (1.9)$$
The pollutant parameter can now be interpreted. Indeed, for pol = 1, model (1.9) becomes
\[
\log \{E(Y|X)|pol = 1\} = \beta_0 + \logspdate(date) + \beta_{11}J_1 + ... + \beta_{16}J_6
+ \beta_2F + \beta_3V + \beta + \sum_{1 \leq i \leq k} \logsp(X_i) + \sum_{1 \leq i \leq j} AR_i.
\] (1.10)

Similarly, for pol = 0, model (1.9) can be written as
\[
\log \{E(Y|X)|pol = 0\} = \beta_0 + \logspdate(date) + \beta_{11}J_1 + ... + \beta_{16}J_6
+ \beta_2F + \beta_3V + \sum_{1 \leq i \leq k} \logsp(X_i) + \sum_{1 \leq i \leq j} AR_i.
\] (1.11)

By subtracting equations (1.10) and (1.11), the following equation is obtained
\[
\log \{E(Y|X)|pol = 1\} - \log \{E(Y|X)|pol = 0\} = \beta,
\] (1.12)
which can also be written as
\[
\log \left\{ \frac{E(Y|X)|pol = 1}{E(Y|X)|pol = 0} \right\} = \beta.
\] (1.13)

The exponential of the parameter \( \beta \) can therefore be interpreted as the relative risk of the variable \( Y \) for an increase of one unity of the value of the pollutant.

Note that another strategy could be adopted. The autocorrelation terms could be inserted in the model before comparing the LOESS function of the pollutant to a linear effect of this pollutant. In this work, it was chosen to add the autocorrelation terms at last. Autocorrelation terms are introduced only in aim to take the serial correlation into account.

1.3.3 Criticism about the use of standard statistical software to fit GAM to epidemiological time series data

Recently major concern was raised about numerical accuracy of the estimates of pollutant effect obtained fitting GAM. Ramsay et al.(2003) and Dominici et al.(2002) identified important critical points in the analysis of epidemiological time series using commercial statistical software which fits GAM by backfitting algorithm. Two criticisms have been established. First, the default convergence criteria of backfitting algorithm defined in S-Plus (and, to a lesser degree, in SAS) are too lax to assure convergence and lead to biased upwards estimates of pollutant effect. Secondly, the estimated standard errors obtained fitting GAM in S-Plus or SAS are biased.

As demonstrated in section 1.3.1, the GAM is a generalization of linear regression. Most of the familiar diagnosis tests for fitted linear regression models have analogues to fitted GAMs. One important exception to this rule is concurvity, the nonparametric analogue of multicollinearity. Multicollinearity is present in the data if some subset of the regressors is highly correlated. It leads to highly unstable and highly correlated parameters estimates associated with the multicollinear variables. Concurvity is a nonparametric extension of this concept. Concurvity occurs when a function \( s(X_p) \) of one of the variables, say \( X_p \), can be approximated by a linear combination of functions \( s(X_1), ..., s(X_{p-1}) \) of the other variables. As is the case for linear regression, the parameter estimates of a fitted GAM are highly unstable if there is concurvity in the data. If data exhibit relevant degree of concurvity, convergence of backfitting algorithm can be very slow (Biggeri et al., 2003). Dominici et al. (2002) showed
that when a smooth function for time and a smooth function for weather are included in the
model, the greater the degree of concavity, the greater is the overestimation of the pollutant
effect.

At present, S-Plus and SAS provides no diagnosis tools for assessing the impact of concur-
vity on a fitted GAM. The inability of the GAMs to detect concavity can lead to misleading
statistical inferences, notably an overstatement of the significance of associations between air
pollution and health status. Because of the way variances are estimated in the S-Plus and
SAS gam functions, the variance estimates do not reflect the instability of the parameter
estimates. The variance estimates produced by the gam function will be biased downwards
in case of concavity. Indeed, the two statistical softwares provide an approximation of the
variance-covariance matrix, which takes into account only the linear component of the variable
that was fit with a smooth function. That leads to confidence intervals on linear parameters
being too narrow and in misleading P-values for hypothesis tests based on variance estimates,
resulting in an inflated type I error.

Some solutions to these problems have been suggested (Samet et al., 2003). It was first
proposed to replace GAM functions with those using stricter convergence criteria. That was
suggested to correct the GAM convergence problem while acknowledging that the problem
with standard error estimates was not addressed. Secondly, it could be interesting to use alter-
native modelling approaches to GAM fitted by backfitting algorithm. The use of generalized
linear models (GLMs) with natural cubic splines, using approximately the same degrees of
freedom as were used in the original GAMs or the use of GAM with penalized regression spline
fitted by the direct method in R software, which correctly computes the variance-covariance
matrix, are two solutions. These solutions were suggested in aim to correct problems with the
standard errors.

In case of the bronchiolitis study, it was decided to evaluate the sensitivity of the results by
fitting: first, GAM by backfitting algorithm using S-Plus with default or stringent convergence
criteria and secondly, GAM by direct method implemented in R 1.6.2 software. A comparison
between S-Plus and SAS results has also been performed.

1.4 Results

The different programs used to obtained the results can be obtained from the authors on
request. An example is presented deeply for S-Plus, and a comparison with R and SAS is
discussed.

1.4.1 Series of number of hospital consultations, results with S-Plus

This section deals with the model building process of the daily number of hospital consult-
tations for infant bronchiolitis for the period between 1997 and 2000, using the Splus method.

For each model tested, a step of graphical validation consisting on examination of the
graph of the residual, and/or, the autocorrelation, and/or, the partial autocorrelation of its
residuals and the graph of the LOESS function for the new introduced variable is performed. 
Due to the huge number of these models, only few representative graphs will be presented
here.
Long term and seasonal variations

Figure 1 displays the daily number of hospital consultations for infant bronchiolitis in the 34 hospitals considered for this study. The series presents clearly some seasonal variations. A peak can be observed each winter.

The autocorrelation of the number of hospital consultations is given on figure 2. The interval around 0 represents the 95% confidence interval of the test of null hypothesis "null autocorrelation". This graph shows that some serial correlation is present in the data. The autocorrelation on the first lag is very high.

![Fig.1 - Daily number of hospital consultations](image)

![Fig.2 - Partial autocorrelation of the daily number of hospital consultations](image)

Fig.3 - Residuals of model 1.14

Fig.4 - Partial autocorrelation of the residuals after introduction of the time.

Long term and seasonal variations are first modelled by introducing the time in the model. A LOESS function is used as smooth function. By looking at the different validation graphs, a span equal to 0.15 is retained. Figures 3 and 4 show respectively the graph of the residuals of the model and the graph of partial autocorrelation of these residuals. After adjustment for the trend and seasonal variations, the partial autocorrelation of the residuals decreases a lot on the first lag and is low on the last lags. Some lags are however still significatively different from 0.

Figure 5 displays the LOESS function corresponding to the time. It represents the part of the logarithm of the number of hospital consultations explained by this covariate. The seasonal variations of the daily number of hospital consultations can clearly been observed. Indeed, the four peaks are present on this graph.

At this step, the model can be written as

\[
\log\{E(nb|X)\} = \beta_0 + \log_{0.15}(date),
\]

(1.14)

with \(nb\) the number of hospital consultations in the 34 hospitals retained for the study.
Short term variations

The next step consists in modelling the short term variations by introducing the variables corresponding to the days of the week (six binary variables $J_1$...$J_6$, with Sunday as reference), official holidays (binary variable : $F$) and holidays (binary variable : $V$). After each introduction, the graphs of partial autocorrelation of the residuals of the model and the part of the logarithm of the number of consultations explained by each covariate were produced and evaluated. The model obtained after this step can then be written as

$$\log\{E(nb|X)\} = \beta_0 + \log_{0.15}(date) + \beta_{11}J_1 + \ldots + \beta_{16}J_6 + \beta_2F + \beta_3V. \quad (1.15)$$

After taking these covariates into account, the partial autocorrelation of the residuals become null on the last lags.

Introduction of the weather factors

All the five weathers factors could be introduced in the model. The minimal temperature is first included in the model. It was decided to choose the lag for this covariate using the "AIC-criterion" and to keep the same lag for the other weather factors. The minimal temperature at lag 1 is retained. Let $tmin1$ this variable. As for the covariate representing the time, a LOESS function is used to model the effect of minimal temperature at lag 1. A span equal to 0.8 is retained by looking at the validation graphs. Then successively, the relative humidity, the daily precipitation, the wind strength and the atmospheric pressure were taken into consideration. As, for the minimal temperature, the examination of the validation graphs allow us to choose the span value. The final model at this step is:

$$\log\{E(nb|X)\} = \beta_0 + \log_{0.15}(date) + \beta_{11}J_1 + \ldots + \beta_{16}J_6 + \beta_2F + \beta_3V + \log_{0.8}(tmin1) + \log_{0.6}(humrel1) + \log_{0.7}(precipit1) + \log_{0.7}(forcvent1) + \log_{0.9}(pressmer1). \quad (1.16)$$

Short terms relations with air pollution

For PM10, 2 indicators were analysed. Let pm10A1 the average of the pollutant concentration over 2 days (the day considered and the previous day) and pm10A5 the average of the pollutant concentration over 6 days (the day considered and the 5 previous days). Only
one of this indicator is included in the model at a time. A LOESS function with a span value equal to 0.7 is used to model the pollutant effect (Le Tertre, 2003). If the introduction of the pollutant is pertinent, the model is compared, as explained in section 1.3.2, to a model presenting a linear effect of this pollutant.

Table 1 displays summary of the p-values corresponding to the introduction of PM10.

**Table 1.1 – Summary of the introduction of the different pollutants**

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>pm10A1</th>
<th>pm10A5</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>0.0626</td>
<td>$1.3736 \times 10^{-6}$</td>
</tr>
</tbody>
</table>

The introduction of the average over 6 days is significant and a new model is obtained:

$$\log\{E(nb|X)\} = A + \log_{0.7}pm10A5,$$

(1.17)

The model containing a linear trend of the pollutant is equivalent to the one including a LOESS function. Table 2 displayed summary of the p-values corresponding to the test of null hypothesis "Model with linear term is equivalent to model with LOESS function".

**Table 1.2 – Summary of the test comparing the model including a linear term and the one with a LOESS function.**

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>pm10A5</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>0.1639</td>
</tr>
</tbody>
</table>

The pollutant for which the linear term is retained is the particle matter less than 10$\mu$m in aerodynamic diameter. These models can then be written as

$$\log\{E(nb|X)\} = A + \beta pm10A5$$

(1.18)

For the particle matter less than 10$\mu$m in aerodynamic diameter, a comparison between the LOESS function and the linear term is shown on figure 7.

![Graph](image.png)

*Fig. 7 - Comparison between LOESS function and linear term for pm10A5.*
A lot of serial correlation remains on the first lags. Some autoregressive terms are therefore introduced in the model. After looking at the validation graphs, it was decided to introduce three autoregressive terms in the model. Hence, the final model have following forms:

$$\log\{E(nb|X)\} = A + \beta pm10A5 + \sum_{1 \leq t \leq 4} AR_t. \quad (1.19)$$

The graph of partial autocorrelation of the residuals of these final models are produced. A graph of partial autocorrelation very close to a white noise can be observed. A graphical comparison between the series and the predictive values for the final model show us that this model is good.

It has been shown, in Section 1.3.2, that the exponential of the parameter $\beta$ can be interpreted as the relative risk of the number of hospital consultations for an increase of one unity of the value of the pollutant. Table 4 displays the estimate of the parameter $\beta$, its standard error, the t-value corresponding, the relative risk for an increase of 10 unitis of the pollutant and its confidence interval.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\beta$</th>
<th>Standard Error</th>
<th>t value</th>
<th>Relative Risk</th>
<th>CI(-)</th>
<th>CI(+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>pm10A5</td>
<td>0.0083</td>
<td>0.0013</td>
<td>6.2506</td>
<td>1.0863</td>
<td>1.0585</td>
<td>1.1149</td>
</tr>
</tbody>
</table>

The model building process was repeated using more stringent convergence parameters than the default setting. Results obtained are exactly the same.

### 1.4.2 Series of number of hospital consultations, results with R

Models obtained in the previous section are rebuilt using R software. LOESS functions are replaced by penalized splines with a number of knots $k$ and a smoothing parameter $\lambda$ instead of a span parameter. The model building process under R is not done from beginning. Concerning the specific days and the weather factors, for an easier comparison with S-Plus results, the covariates retained in section 1.4.1 is used here. Hence, the five weather factors will be used at lag 1. Moreover, only one pollutant, the particle matter less than 10$\mu$m in aerodynamic diameter, for which the relation hospital consultations-pollutant has been expressed in relative risk (in section 1.4.1) is considered.

A model without any pollutant is first built. Spline functions are used to model the effect of the time and of the weather factors. By looking at the validation graphs, the selected model is the following:

$$\log\{E(nb|X)\} = \beta_0 + s_{40\&5.8\times10^4}(date) + \beta_1 J_1 + \ldots + \beta_{16} J_6 + \beta_2 F + \beta_3 V + s_{5\&5.1\times10^{11}}(tmin1) + s_{8\&5.6\times10^{14}}(humidity1) + s_{6\&2.3\times10^{5}}(precipit1) + s_{5\&1\times10^{6}}(forcevent1) + s_{5\&8\times10^{6}}(pressmer1),\quad (1.20)$$

with $s_{40\&5.8\times10^4}(date)$ a penalized spline of the time using a number of knots $k$ equal to 40 and a smoothing parameter $\lambda$ equal to $5.8 \times 10^4$. The graph of the partial autocorrelation of the residuals of this model was produced and examined. Only for very few lags, the partial autocorrelation is significatively different from 0.
The next step consists in the introduction of the pollutant. The particle matter less than 10\(\mu m\) improve significantly (p-values respectively equal to \(3.6085 \times 10^{-5}\)) the model.

The penalized splines corresponding to this pollutant was examined. The number of hospital consultations increases with the pollutants. Since this function is almost perfect line, model including linear term for the pollutant is preferred to the one including non parametric function of the same pollutant. This new model is given below (1.21).

\[
\log\{E(nb|X)\} = \beta_0 + s_{40\&5.8\times10^4(date)} + \beta_{11}J_1 + \ldots + \beta_{16}J_6 + \beta_2F
\]
\[
\phantom{\log\{E(nb|X)\}} = \beta_3 V + s_{5\&5.1\times10^4(tmp1)} + s_{8\&5.6\times10^4(humrel1)} + s_{6\&2.3\times10^6(\text{precip1})}
\]
\[
\phantom{\log\{E(nb|X)\}} = s_{5\&1\times10^6(\text{forcvent1})} + s_{5\&8\times10^6(\text{pressmer1})} + \beta_{pm10.45}
\]

(1.21)

The graph of the partial autocorrelation of the residuals of the final model was produced. Almost a white noise can be observed. Therefore, it was decided to do not introduce autoregressive terms in this case.

Again, the exponential of the parameter \(\beta\) can be interpreted as the relative risk of the number of consultations for an increase of one unity of the value of the pollutant. Table 7 displays the estimate of the parameter \(\beta\), its standard error, the t-value corresponding, the relative risk for an increase of 10 unities of the pollutant and its confidence interval.

<table>
<thead>
<tr>
<th>Model</th>
<th>(\beta)</th>
<th>Standard Error</th>
<th>t value</th>
<th>Relative Risk</th>
<th>CI(-)</th>
<th>CI(+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(pm10\ (1.21))</td>
<td>0.0043</td>
<td>0.0010</td>
<td>4.159</td>
<td>1.0441</td>
<td>1.0231</td>
<td>1.0655</td>
</tr>
</tbody>
</table>

Using p-splines with R, the estimate of the increase in consultations dropped to 4.4 per cent per 10mg/m3 increase in PM10 with a standard error of 0.0010, compared with 8.6 per cent (se : 0.0013) using loess with S-Plus.

1.4.3 SAS results

As in S-Plus, proc GAM in SAS allows to fit GAMs using LOESS functions. However, the smoothing parameter differs from S-Plus. Instead of a "span parameter", the LOESS function needs a "DF parameter". An option allows to ask the value of the smoothing parameter to be selected by generalized cross validation. The default method in SAS uses \(DF = 4\) for each smooth function. It is also permitted to choose a particular \(DF\) for each function. This last method is equivalent to play with the spans in S-Plus.

In aim to compare S-Plus and SAS software, the models obtained using S-Plus have also been implemented in SAS. They are first programmed using, for each LOESS function, the \(DF\) value given in the S-Plus output. The LOESS functions corresponding to the different covariates of the model can be obtained. They look, more or less, to the ones obtained using S-Plus.

The use of SAS software to apply this method on the bronchiolitis data has some disadvantages. First, SAS looks for functions with values of \(DF\) as close as possible to the values asked. That leads to smooth functions different from those asked. For instance, the obtained graph does not show, as it is expected, the four peaks corresponding to the seasonal variations of the
daily number of hospital consultations. Secondly, it takes longer than using S-Plus. Indeed, for the bronchiolitis data, a program runs several hours in SAS while it is almost instantaneous in S-Plus. Finally, the validations graphs used to construct the model step by step are not provided in the options of the GAM-procedure in SAS. They can be obtained more easily in S-Plus.

An interesting challenge could be building a sas macro that allows with sas procedure (gam and others, like Arima and Graphs) to performs well Schwartz method.

1.5 Discussion

Numerous time series studies have linked levels of particulate air pollution to daily mortality and cardio-respiratory hospitalisations. However, the study of the specific role of air pollution as risk factor for specific respiratory disease is not usual. This study, conducted in Paris during four following winters, found that prevailing levels of PM10 had measurable short term effects on hospital consultations for an infectious disease in infant, bronchiolitis.

Generalized additive models (GAMs) with smoothing splines were used in this purpose and models are presented in details in this paper. Because of the critical points identified in the analysis of epidemiological time-series using commercial statistical software which fits GAMs by backfitting algorithm, a sensibility analysis was performed using S-Plus, R and SAS softwares.

The model building process was done in accordance with the method proposed by J Schwartz, controlling for weather, season and other longer-term time-varying factors to minimize confounding of the effect estimates for the air pollutant. Models were first constructed using LOESS functions in the gam function S-Plus. PM10 (particle matter less than 10µm) present an effect that can be expressed in term of a relative risk of hospital consultation for an increase of a certain amount of the pollutant concentration. An increase of 10µg/m$^3$ of the particle matter less than 10µm raises the risk of hospital consultation for bronchiolitis of 8.6% [95%CI.[5.9;11.5]]. Because it was raised that the default convergence criteria of backfitting algorithm would be too lax to assure convergence, the model building process was repeated using more stringent convergence parameters than the default setting; results obtained were identical.

The models obtained in S-Plus were then rebuilt using R software. LOESS functions were replaced by penalized splines. An increase of 10µg/m$^3$ of the particle matter less than 10µm raises this risk of only 4.4% [95%CI.[2.3;6.6]]. S-Plus provides no diagnosis tools for assessing the impact of concurrence on a fitted GAM. S-Plus version 3.4 leded to an overstatement of the significance of associations between air pollution and health status. Moreover, S-Plus provides an approximation of the variance-covariance matrix, which takes into account only the linear component of the variable fitted with a smooth function. That would lead to confidence intervals on linear parameters being too narrow. As it was expected, the relatives risks obtained using S-Plus are higher than the ones obtained with R. However, with this data set, confidence intervals around the estimate are similar (and even a little larger) in S-Plus than in R.

The model building process is easier to implement in S-Plus than in R. Indeed, using LOESS function in S-Plus, values must be chosen for the span of each covariate for which the effect is modelled using a smooth function. Using penalized splines in R, values for two parameters ($k$ and $\lambda$) have to be selected for each covariate. Therefore, for ease of comparison, we build first the models using S-Plus before working with R.

Using GAM in SAS is applicable with more difficulty. Indeed, the validation graphs used
to construct the model step by step are not provided in the options of the GAM-procedure in SAS. Create SAS macros to allow applying more easily the Schwartz method could be done. However, in case of the bronchiolitis data, SAS software seems to be not very appropriate because of the time necessary to run the different programs. Moreover, implementation in SAS presents the same problems than in S-Plus.

The methodologic issues in time-series analysis of air pollution epidemiology are important as the air pollution effect is small and possibly confounded by varying factors which are correlated with pollution exposures. The potential for incorrect standard errors in GAMs was known (référence 4), but the new methods had not been implemented in widely used software, and so have not reached epidemiologists. Actually, the S-Plus default convergence parameters have already been revised in the new S-Plus version and revisions of GAM software implementation, allowing "exact" calculations of the standard errors are underway.

Our work has permitted to show that the pollutants would have a short-term incidence on infant bronchiolitis and to confirm the overestimation of the risks by the implementation of GAMs in S-Plus, version 3.4.

References


- This study was supported by Conseil Scientifique de l’ANTADIR. - The authors are grateful to Délégation à l’Information Médicale et à L’Épidémiologie, AP-HP, Paris who provides ERBUS data.