

BIBLIOTECA DI SCIENZE STATISTICHE
BID. PUV0996797
ACQ. 84/105 INV. 85318
Collocazione 5-666 WP. 13/2001

**MORTALITY AND AIR
POLLUTION IN PHILADELPHIA:
A DYNAMIC GENERALIZED
LINEAR MODELLING
APPROACH**

M. Chiogna, C. Gaetan

2001.13

Università di Padova
BIBLIOTECA DI SCIENZE STATISTICHE
Via Cesare Battisti, 241 - 35121 PADOVA

Dipartimento di Scienze Statistiche
Università degli Studi
Via C. Battisti 241-243
35121 Padova

Ottobre 2001

BIBLIOTECA DI SCIENZE STATISTICHE

BID. 100.100.100

ACC. 100.100.100 INV. 100.100.100

Collocazione 100.100.100

Università di Padova
BIBLIOTECA DI SCIENZE STATISTICHE
Via Cesare Battisti, 241 - 35121 PADOVA

Il presente documento è
depositato nella
Biblioteca di Scienze Statistiche
dell'Università di Padova

Mortality and Air Pollution in Philadelphia: a Dynamic Generalized Linear Modelling Approach

Mortalità ed Inquinamento Atmosferico a Philadelphia: un Approccio secondo i Modelli Dinamici Lineari Generalizzati

Monica Chiogna and Carlo Gaetan

Dipartimento di Scienze Statistiche, Via Cesare Battisti, 241-243, Padova, Italy

e-mail: [monica,gaetan]@stat.unipd.it

Riassunto: Con riferimento allo studio degli effetti a breve termine dell'esposizione ad inquinamento atmosferico, si propone l'utilizzo dei modelli lineari generalizzati dinamici. Tale classe di modelli appare infatti idonea per gestire il complesso meccanismo di dipendenza che caratterizza i fenomeni oggetto di studio. La metodologia viene applicata allo studio della mortalità per cause non accidentali a Philadelphia (1974–1988). I risultati evidenziano una buona flessibilità dello strumento di modellazione, consentendo in particolare l'utilizzo di un ridotto numero di variabili esplicative.

Keywords: Epidemiologic Time Series, Dynamic Models.

1. Introduction and motivation

Various applied fields, like environmental statistics or environmental epidemiology, deal with time series data in the form of discrete or non-normal outcomes. In environmental epidemiology, a key problem is the role of serial correlation in the modelling framework. Serial correlation on the response is due to different sources of causes. First of all, outcomes depend on serially correlated explanatory variables, so that the time series structure of the covariates imparts a highly structured pattern of interdependence on the response. Then, the effect of the explanatory variables on the outcomes usually lasts some time; for example, the effect of a high pollution event on population health spreads over some days, although the effect's mechanisms is unfortunately unknown. This, again, depends partly on the serial correlation of the pollutants, partly on the natural response mechanism of the human body to exposure to toxic agents.

A natural way to deal with such issues would be to develop association models in which the dependence structure within the explanatory variables and between covariates and response is correctly accounted for. However, in most cases neither the dependence mechanism on the explanatory variables nor an adequate knowledge of the physical association between response and covariates is known, so that an explicit probabilistic model of association is rarely available.

In the environmental epidemiology literature, the usual modelling strategy (see Brumback *et al.* (2000), for up to date references) is based on estimating proper

Generalized Linear Model (GLM) or Generalized Additive Models (GAM) with the assumption of independent outcomes. The temporal behavior is controlled by mean of a cyclic function inserted in the regressor term. In the diagnostic phase, checks for the presence of serial correlation in the residuals are performed. In general, residuals' serial correlation happens when the model for the temporal component is inadequate to pick up all the fluctuations of the underlying outcome behavior. In these cases, a more accurate modeling of the temporal component can solve the problem. If this is not the case, the modeling strategy can be extended to account for serially correlated terms.

In the literature, two main approaches can be followed to add autocorrelation to the standard GLM or GAM setting: either a latent autocorrelated time series error is assumed for the model (generalized linear/additive model with time series error), which means that correlation between two subsequent outcomes is a known function of the marginal means of the outcomes and perhaps of some additional parameters, or correlation is inserted into the model by making the current outcome explicitly depend on past outcomes (transitional generalized linear/additive model).

Various difficulties are related to these extensions, like computational difficulties, model checking, etc. In this paper we propose to employ a different modelling strategy based on Dynamic Generalized Linear Models (DGLMs) (Fahrmeir and Tutz (1994)), which extends the first above mentioned approach. We apply this modelling framework to the analysis of daily death counts in Philadelphia (Kim *et al.* (1999)). The counts are modelled by a Poisson distribution having mean driven by a latent Markov process. In this setting, serial dependence is added to the model structure by making use of random coefficients supplemented by prior distributions adequate to take autocorrelation into account, such as for example random walks. This implies that time-varying covariates are allowed to enter the model not only via the conditional mean but also linked to the latent process. In our view, this framework allows to more neatly explore the dependence structure of the data. Estimation is performed by extended Kalman filter and smoother.

The outline of this paper is as follows. In Section 2 we introduce some features of our modelling approach, sketching also the inference procedure. Section 3 briefly describes the Philadelphia data set. Finally, Section 4 contains the results and some concluding remarks.

2. Dynamic Generalized Linear Models

Consider a series of counts $\{Y_t\}$ recorded at equally spaced times $t = 1, \dots, T$ along with a vector of covariates $\mathbf{x}_t \in R^p$. Assume that it is reasonable to divide the vector of covariates into two components $\mathbf{x}_t = (\mathbf{x}'_{1t}, \mathbf{x}'_{2t})'$, where the first component, \mathbf{x}_{1t} , includes covariates which we expect to contribute to the 'nucleus' of the underlying mean tendency of the counts and the second component, \mathbf{x}_{2t} , includes perturbing factors, whose influence can be thought of as being constant over time. If the central tendency of the counts may be thought of as resulting from the influence of both types of covariates, then it is reasonable to model the effect of \mathbf{x}_{1t} by a univariate latent process $\phi_t = \phi_t(\mathbf{x}_{1t})$ and to fix the effects of \mathbf{x}_{2t} . Therefore, we assume that

the conditional distribution of Y_t given ϕ_t follows a Poisson distribution

$$Y_t|\phi_t \sim \mathcal{P}(\exp\{\phi_t + \mathbf{x}_{2t}'\boldsymbol{\gamma}\})$$

with $\boldsymbol{\gamma}$ representing the fixed regression coefficients for \mathbf{x}_{2t} . For the latent process ϕ_t we consider the following specification:

$$\begin{aligned}\phi_t &= (\omega_t + \mathbf{x}_{1t}'\boldsymbol{\beta}_t) \\ \omega_t &= 2\omega_{t-1} - \omega_{t-2} + \delta_t \quad \text{with} \quad \delta_t \sim N(\mu_\delta, \sigma_\delta^2). \\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\xi}_t \quad \text{with} \quad \boldsymbol{\xi}_t \sim N(\boldsymbol{\mu}_\xi, \Sigma_\xi).\end{aligned}$$

In this formulation ω_t is a discrete-time analog of a continuous-time cubic spline so that the model appears as a semi-parametric model. Also the coefficients of the long-term explanatory variables move dynamically in time. This is a crucial aspect of the model. Firstly, it is possible to grasp long term changes in the effects of the covariates. Secondly, the dynamic on the coefficients allows to capture the “carry-over” effect of the covariates which usually causes the effect at time t to be influenced by covariates at previous times. The previous setting falls into the general framework of Dynamic Generalized Linear Model (DGLM) (Fahrmeir and Tutz (1994)). The latent process ϕ_t and the parameters $\boldsymbol{\gamma}$ can be cast in the transition model

$$\boldsymbol{\alpha}_t = F\boldsymbol{\alpha}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{1}$$

with $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, Q)$, $\boldsymbol{\alpha}_0 \sim \mathcal{N}(\mathbf{a}_0, P_0)$ and \mathbf{a}_0, P_0, Q, F hyperparameters. The observation model is given by

$$Y_t|\phi_t \sim \mathcal{P}(\exp\{\mathbf{z}_t'\boldsymbol{\alpha}_t\}). \tag{2}$$

To exemplify, assume for convenience that $\mathbf{x}_t = (x_{1t}, x_{2t})'$. Then,

$$\boldsymbol{\alpha}_t = \begin{pmatrix} \omega_t \\ \omega_{t-1} \\ \boldsymbol{\beta}_t \\ \boldsymbol{\gamma} \end{pmatrix} = \begin{pmatrix} 2 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \omega_{t-1} \\ \omega_{t-2} \\ \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\gamma} \end{pmatrix} + \begin{pmatrix} \delta_t \\ 0 \\ \boldsymbol{\xi}_t \\ 0 \end{pmatrix} = F\boldsymbol{\alpha}_{t-1} + \boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \text{diag}(\sigma_\delta^2, 0, \sigma_\xi^2, 0))$ and $\mathbf{z}_t = (1, 0, x_{1t}, x_{2t})'$.

2.1 Moment structure

We will explore the nature of the serial correlation implied on the counts by the dynamical structure. Consider first the mean, \mathbf{a}_t , the variance, P_t , and the autocovariance function, $P_{t,t+h}$, of $\boldsymbol{\alpha}_t$. By recursion, it is possible to show that:

$$\mathbf{a}_t = F^t \mathbf{a}_0,$$

with $F^t = F \cdots F$, t times and $F^0 = I$. Analogously:

$$P_t = \sum_{k=0}^{t-1} F^k Q F'^k + F^k P_0 F'^k$$

and

$$\text{Cov}(\alpha_{t+h}, \alpha_t) = F^h P_t.$$

Therefore, $\alpha_t \sim \mathcal{N}(\mathbf{a}_t, P_t)$, which yields for the latent process:

$$\mathbf{z}'_t \alpha_t \sim \mathcal{N}(\mathbf{z}'_t \mathbf{a}_t, \mathbf{z}'_t P_t \mathbf{z}_t)$$

and

$$\text{Cov}(\mathbf{z}'_{t+h} \alpha_{t+h}, \mathbf{z}'_t \alpha_t) = \mathbf{z}'_{t+h} F^h P_t \mathbf{z}_t.$$

We now turn to the moment structure of the observed counts, for which we have:

$$\mu_t = E(E(Y_t | \alpha_t)) = E(\exp(\mathbf{z}'_t \alpha_t)),$$

$$\sigma_t^2 = E(\text{Var}(Y_t | \alpha_t)) + \text{Var}(E(Y_t | \alpha_t)) = E(\exp(\mathbf{z}'_t \alpha_t)) + \text{Var}(\exp(\mathbf{z}'_t \alpha_t)).$$

This yields

$$\mu_t = \exp(\mathbf{z}'_t \mathbf{a}_t + \mathbf{z}'_t P_t \mathbf{z}_t / 2)$$

$$\sigma_t^2 = \mu_t^2 (\exp(\mathbf{z}'_t P_t \mathbf{z}_t) - 1)$$

$$\text{Cov}(Y_{t+h}, Y_t) = \mu_{t+h} \mu_t (\exp(\mathbf{z}'_{t+h} F^h P_t \mathbf{z}_t) - 1),$$

from which we derive that:

$$\text{Corr}(Y_{t+h}, Y_t) = \frac{\exp(\mathbf{z}'_{t+h} F^h P_t \mathbf{z}_t) - 1}{\sqrt{(\exp(\mathbf{z}'_{t+h} P_{t+h} \mathbf{z}_{t+h}) - 1)(\exp(\mathbf{z}'_t P_t \mathbf{z}_t) - 1)}}.$$

Previous results show that this model setting allows to incorporate non stationary second-order processes.

2.2 Inference

DGLMs have two unknown quantities: the state vector α_t and the hyperparameters. We summarize the hyperparameters in the vector λ , we assume for the moment λ fixed and known and we are interested in dealing first with α_t . Let $\alpha_t^* = (\alpha'_0, \dots, \alpha'_t)'$ and $Y_t^* = (Y_1, \dots, Y_t)'$, the conditional distribution of α_T^* given the observations

$$p(\alpha_T^* | Y_T^*) \propto \prod_{t=1}^T p(Y_t | \alpha_t) \prod_{t=1}^T p(\alpha_t | \alpha_{t-1}) p(\alpha_0) \quad (3)$$

is non-normal. Note that in this case the conditional means and the conditional modes are not equivalent. Due the complicated form of the conditional distributions involved, inference requires some approximation. Simulation based estimation, in particular MCMC methods, has been proposed for dealing with this problem (see Shephard and Pitt (1997), among others). In this paper we have chosen the approach proposed by Fahrmeir and Tutz (1994) and we consider (3) as a penalized likelihood

avoiding a Bayesian interpretation. More precisely, taking the logarithm of the conditional density, $PL(\boldsymbol{\alpha}_T^*)$, this yields to

$$PL(\boldsymbol{\alpha}_T^*) = \text{const} + \sum_{t=1}^T l(\boldsymbol{\alpha}_t) - \frac{1}{2}(\boldsymbol{\alpha}_0 - \mathbf{a}_0)' P_0^{-1}(\boldsymbol{\alpha}_0 - \mathbf{a}_0) - \frac{1}{2} \sum_{t=1}^T (\boldsymbol{\alpha}_t - F\boldsymbol{\alpha}_{t-1})' Q^{-1}(\boldsymbol{\alpha}_t - F\boldsymbol{\alpha}_{t-1}). \quad (4)$$

with $l(\boldsymbol{\alpha}_t) = \log p(Y_t | \boldsymbol{\alpha}_t)$. The function (4) is a penalized log-likelihood criterion so the conditional modes $\hat{\boldsymbol{\alpha}}_T^* = \text{argmax}_{\boldsymbol{\alpha}_T^*} p(\boldsymbol{\alpha}_T^* | Y_T^*)$, are maximizers of $PL(\boldsymbol{\alpha}_T^*)$. Algorithmic solutions can be efficiently obtained by iterative Kalman filtering and smoothing (Fahrmeir and Wagenpfeil (1997)). The hyperparameters $\boldsymbol{\lambda}$ can be interpreted as smoothing parameters and these can be selected as minimizers of a generalized cross-validation criterion.

Diagnostics deserve some care. In particular Pearson-type residuals loose the usual properties. For more appropriate inference we use the P-scores as suggested by Früwirth-Schnatter (1996), and we perform standard diagnostic tools for generalized linear models as well as dynamic linear models for detecting model's inadequacies, if any.

3. Data

The first substantial analysis of a U.S. epidemiological time series data set was by Schwartz and Dockery (1992), using data from Philadelphia and it was quickly followed by a number of studies of other cities. The main Philadelphia data set used by the researchers consisted of 14 years of daily deaths data (1974–1988) with associated measurements of temperature and dewpoint, i.e. the two meteorological variables which are believed to be the most important confounders, and five pollutants: total suspended particulate (TSP), sulphur dioxide (SO₂), nitrogen dioxide (NO₂), carbon monoxide (CO) and ozone (O₃).

In the present analysis, we have used the same data set as in Kim *et al.* (1999), and we have focused on the analysis of TSP effects on deaths in the population aged 65 and over, since this is the group most at risk. Table 1 includes summary statistics for the daily air pollutant concentration, key meteorology and death counts recorded during the study period.

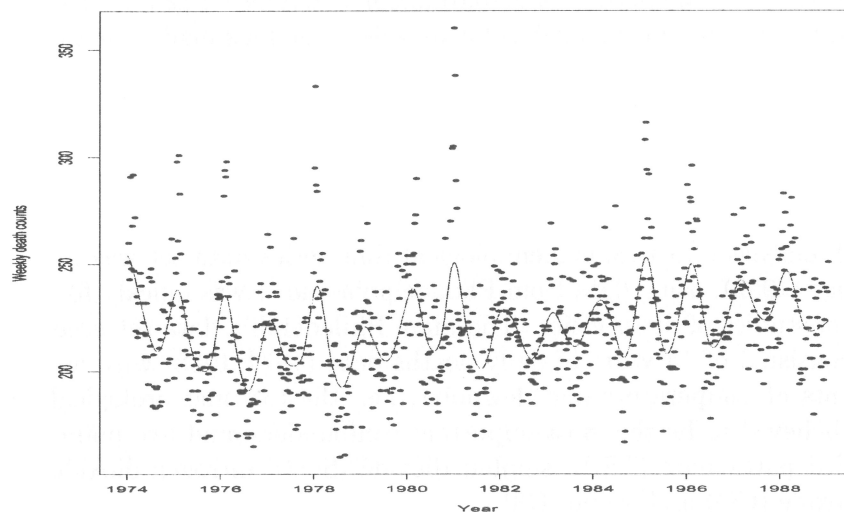
Table 1: *Summary statistics for Philadelphia: Mortality age 65+, temperature and dew-point (°F), TSP (µg/m³).*

Var.	Mean	SD	25%	50%	75%
Mort	31.5	6.4	26	31	38
Temp	54.3	17.8	40.0	55.3	70.3
Dew	42.3	19.1	27.8	43.5	58.8
TSP	67.3	26.9	47.5	63.0	72.0

Figure 1, which plots weekly deaths counts along with a non parametric estimate of the temporal behaviour, shows that there is an irregular seasonal effect, which

cannot be explained solely through the dependence of deaths on either meteorology or pollution and need to be properly modelled. Moreover, extensive preliminary analyses show that deaths decrease against both temperature and dewpoint until a threshold value is reached, around 75° F for temperature and 60° F for dewpoint, after which deaths start to increase with increasing temperature or dewpoint. Similar analyses for TSP show a general increase in deaths with pollutant, although it is questionable whether there is any real effect below about 100 $\mu\text{g}/\text{m}^3$. To develop the models, we take advantage of the exploratory analysis and of previous studies (see Kim *et al.* (1999)) to construct and select sound covariates, although we do not aim at building models which are strictly comparable with the models already published.

Figure 1: Weekly deaths counts for the years of study (1974–1988) along with a non parametric estimate of the temporal behaviour (solid line). Dotted vertical lines mark the beginning of each calendar year.



4. Results and conclusions

Our model building strategy started from the simplest models, i.e. models including the pollutant and the most relevant meteorological variables. If variables resulted to be not significant, i.e. the corresponding time-varying confidence bands never included zero, they were removed from the models.

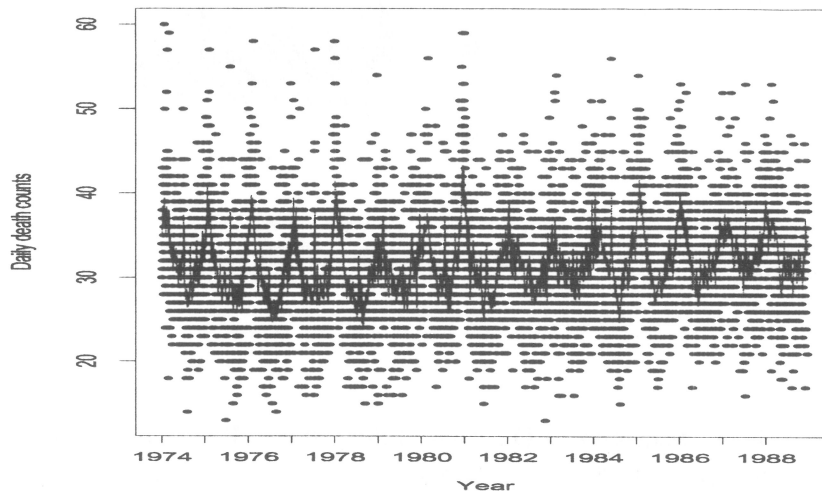
As we assumed that the counts reflected an underlying tendency of the severe air pollution events, combined with adverse meteorological conditions, to cause non-accidental death, we included in the \mathbf{x}_{1t} component vector those covariates which measured exposure to pollution and meteorological conditions.

Long term trends and seasonal fluctuations were controlled by making use of a discrete version of a spline component modelled by a second order random walk.

Controlling for weather was achieved by including a time varying coefficient for the mean temperature of the previous 3 days less than 80°F and a fixed effect for temperature above the same cutoff. Moreover, a time varying effect for the mean dew point temperature of the previous 3 days was also introduced.

All analyses were carried out using R functions and C routines written by ourselves. Figure 2 shows the time series of the observed counts along with the fitted values.

Figure 2: *Time series of the observed counts along with fitted values.*

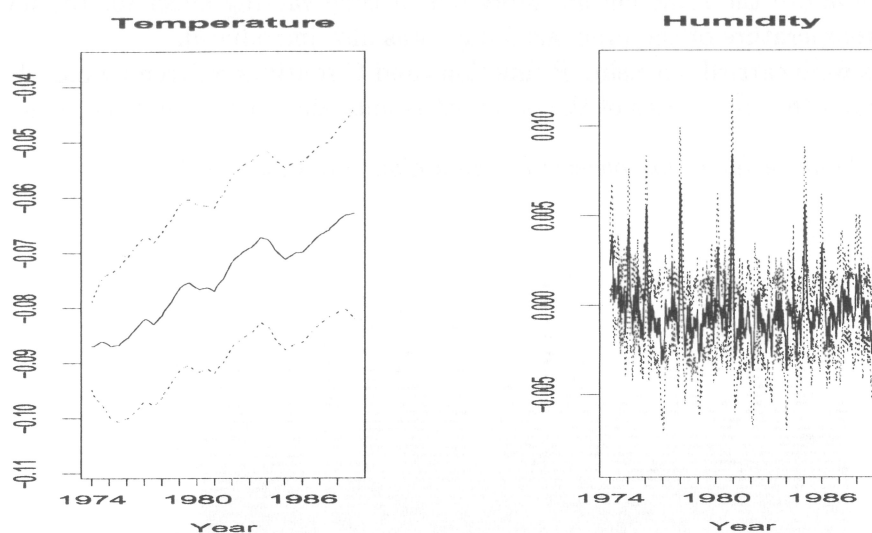


Results showed a significant and positive effect of TSP on health. The pollutant's effect resulted to be fixed, i.e. the estimated coefficient variance was near to zero, meaning that the three day lagged measure was sufficient to capture the carry-over effect of the pollutant.

Figure 3 shows the estimated trajectories for the coefficients of the temperature below 80°F and of the dew point temperature and highlights the time varying effects of the variables. The temperature effect appears to be significant over the study period, whereas controversial appears the interpretation of the dew point temperature. It appears that most of the dynamic behaviour is captured by humidity so that it becomes difficult to assess significance of the variable and the strength of its effect.

The aim of our work was to explore the extent to which the DGLM setting allows to improve modelling procedures for epidemiological time series studies, where a complex dependence structure among variables is observed. Results seem satisfying in terms of flexibility of the modelling approach and of parsimony, where parsimony is to be intended with respect to the number of explanatory and confounding variables that need to be included in the model and not with respect to the number of parameters to be estimated.

Figure 3: *Estimated trajectories for the coefficients of temperature below 80 °F and of the dew point temperature with point-wise 95% confidence intervals.*



References

- Brumback, B.A., Ryan, L.M., Schwartz, J.D., Neas, L.M., Stark, P.C., Burge, H. (2000) Transitional regression models with application to environmental time series. *Journal of the American Statistical Association*, **95**, 16–27.
- Fahrmeir, L., Tutz, G. (1994) *Multivariate Statistical Modelling Based on Generalized Linear Models* Springer-Verlag, New York.
- Fahrmeir, L. and Wagenpfeil, S. (1997) Penalized likelihood estimation and iterative Kalman smoothing for non-gaussian dynamic regression models. *Computational Statistics and Data Analysis*, **24**, 295–320.
- Früwirth-Schnatter, S. (1996). Recursive residuals and model diagnostics for normal and non-normal state space models. *Environmental and Ecological Statistics*, **3**, 291–309.
- Kim Y., Spitzner D., Zhang Z., Smith R.L., Fuentes M. (1999) Accounting for multiple pollutants in pollution-mortality studies. In *Proceedings of the ASA – Biometrics Section*, 1–10.
- Schwartz, J. and Dockery, D.W. (1992). Increased mortality in Philadelphia associated with daily air pollution concentrations. *A. Rev. Respr. Dis.*, **145**, 600–604.
- Shephard, N. and Pitt, M.K. (1997). Likelihood analysis of non-gaussian measurement time series. *Biometrika*, **84**, 653–667.