Mortality and air pollution: a dynamic generalized linear modelling approach

Monica Chiogna and Carlo Gaetan
Università di Padova
[monica,gaetan]@stat.unipd.it

Genova, June 18-22, 2002

Background

Epidemiological evidence is gaining increasing relevance in the establishment of national air quality standards (see U.S Federal register, 1997).

Evidence comes from time series regressions of mortality/morbidity counts on various covariates (pollution, meteorology, confounders, etc.) . Focus is on measuring the risk of negative health outcomes.

Only recently has the statistical literature started to evaluate the impact of the methodology on the assessment of the risk.

A recent example.....
(June 2002)

Important New Findings from NMMAPS Investigators. 

The investigators of HEI's National Morbidity, Mortality, and Air Pollution Study (NMMAPS) at Johns Hopkins University Bloomberg School of Public Health have reported important new findings about the statistical techniques used to conduct their studies (and some other studies like theirs at other research centers). HEI has published a letter describing these findings and HEI's plans to carefully review new results emerging from the NMMAPS study. HEI plans to periodically update this site as further understanding of these findings emerge.

`http://www.healtheffects.org/`

Usual strategy

Based on building proper GLMs or GAMs in three steps:

1. adjust for temporal confounding in the response

2. adjust for meteorological confounding

3. insert pollutant(s)

Finally, a check on residual autocorrelation is performed. If autocorrelation is detected, possible solutions are:

- Observation-driven models (Zeger and Qaqish (1988), Baccini *et al.* (1999), Brumback *et al.* (2000))

  e.g. $g(\mu_t) = \alpha y_{t-k} + \beta x_t$

- Parameter-driven models (Zeger (1988), Fahrmeir and Tutz (1994), Jørgensen *et al.* (2000))

e.g. $g(\mu_t) = \beta x_t + \delta_t$ with δ_t random

General criticisms

- Associations are the spurious result of imperfect control for seasonality and longer term tends
- Associations are spurious results of confounding by other pollutants which do effect mortality
- Associations are not valid because use ambient not personal exposure levels ("measurement error")
- Associations are of little public health importance since only frail persons are affected by pollution ("harvesting")

Let us start from the beginning

Key feature: time series structure of the data.
It depends on:

1. secular trends and seasonal variation in the response;

2. long–term trends and seasonal variations in the covariates;

3. carry–over effects;

and it gives rise to

1. confounding;

2. multicollinearity;

3. serial correlation.

A unified framework

- Proposal:

  To tackle the complex modelling task by making use of Dynamic Generalized Linear Models (Fahrmeir and Tutz, 1994).

- How?

  We consider regression models based on state space models for the outcomes; temporal features are added to the model structure by making use of random coefficients supplemented by prior processes, such as e random walks, that can handle autocorrelation.

In the following

Outline of the talk

1. Description of the Birmingham (Alabama) case study.

2. Dynamic generalized linear models.

3. Results and perspectives.

Case Study

Daily mortality in Birmingham, Alabama 1985–1988

- Response:
  - deaths from non accidental causes (people aged 65 and over)

- Covariates
  - air pollutant: PM_{10}
  - metereology: temperature (max, min), specific humidity, dew point temperature

A look to the data

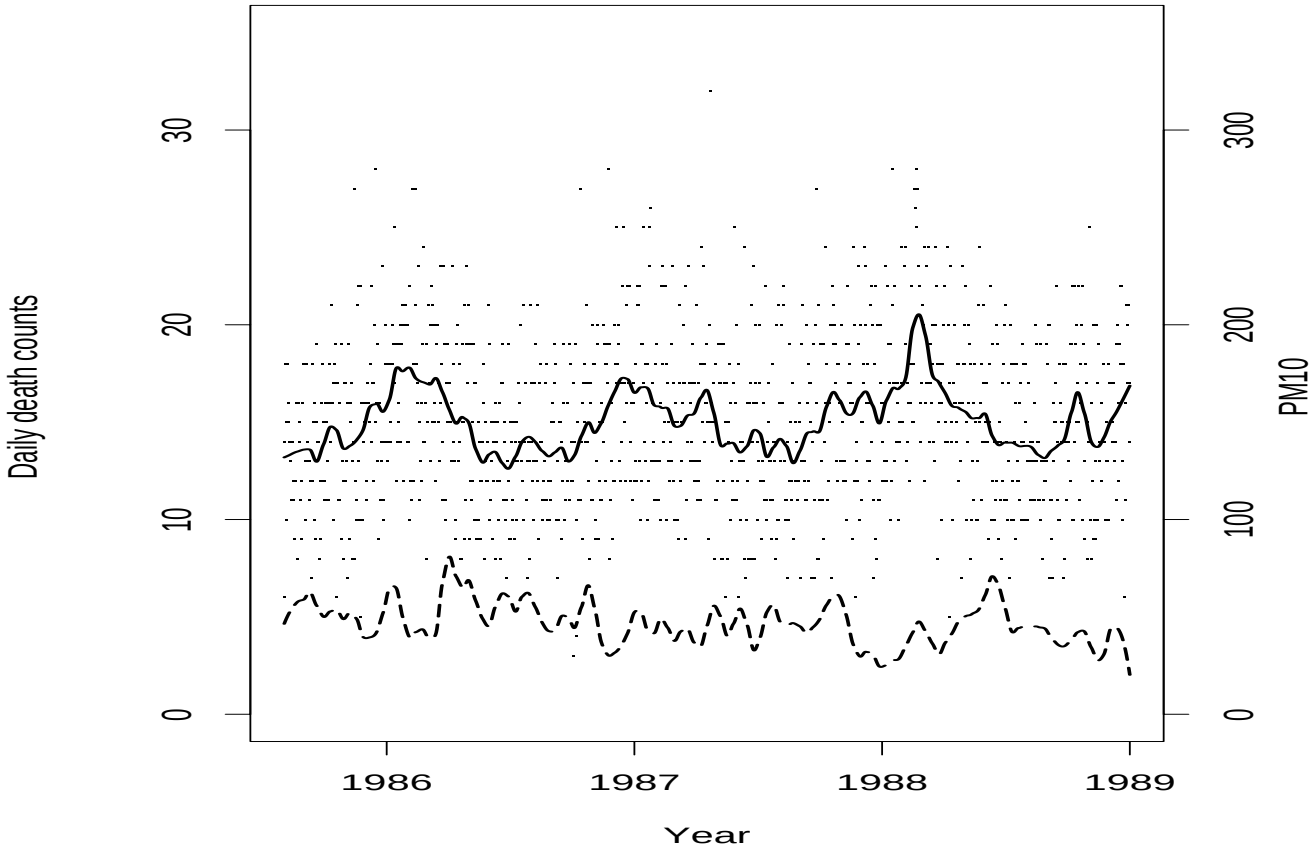

Figure 1: Time series of the observed daily death counts along with the `loess` smoother of the counts (solid line) and of the PM10 values (dotted line).

History

- Schwartz (1993) : significant positive effect
- Samet $et\ al.$ (1995): confirmed Schwartz's results
- Roth and Li (1996): 2400 models (!!) of which 4 showed significant positive effect and 3 significant negative effect (period of analysis 1986–1990)
- Clyde (2000): GLM + Bayesian Model Averaging leads to lower estimates of risk
- Smith $et\ al.$ (2000): controversial results

Models in Smith et al. (2000)

Different models with different adjustments for metereology and exposure.

Different modelling strategies (LM, GLM, GAM).

Various lagged effects and distributed lagged effects.

Careful diagnostics on overdispersion, serial correlation and lacks of fit.

Some problems raised by Smith et al.

Results depend mostly more on the use of covariates than on the modelling strategy.

- Definition of exposure measure and its lagged effects:

lag	effect	significance
0	< 0	yes
1	> 0	yes
2	> 0	no
3	> 0	no
0–2	> 0	yes
1–3	> 0	yes

- Definition of the lagged effects for metereology.

- Possible time trend in the PM_{10} coefficient?

Dynamic generalized linear models

$$y_t|\phi_t \sim Po(\exp\{x_t^T\alpha + \phi_t\})$$

with:

$\phi_t \rightarrow$ latent process (even non stationary) describing the dynamic of the central tendency.

$x_t \rightarrow$ "short-term" covariates: modulating factors acting on the mean

Possible specifications for ϕ_t

1. long term trend

   $\phi_t = \omega_t$

   $\omega_t = 2\omega_{t-1} - \omega_{t-2} + \delta_t \qquad$ with $\delta_t \sim N(\mu_\delta, \sigma^2_\delta)$

   discrete version of a spline for a trend component

2. long term trend + long term covariates expected to have impact on the response

   $\phi_t = (\omega_t + b_t x_t)$

   $b_t = b_{t-1} + \xi_t \qquad$ with $\xi_t \sim N(\mu_\xi, \sigma^2_\xi)$

NB: having a dynamic on the coefficients of the explanatory variables allows to capture the "carry-over" effect $\Rightarrow$ no need to include lagged values

General setup

Let

$$\begin{aligned} & x_t^T\alpha + \phi_t = Z_t\theta_t \\ y_t|\theta_t \quad &\sim \quad p(y_t|\theta_t) = Po(\exp Z_t\theta_t) \\ \theta_t|\theta_{t-1} \quad &\sim \quad p(\theta_t|\theta_{t-1}) = \mathcal{N}(F\theta_{t-1}, Q) \\ \theta_0 \quad &\sim \quad p(\theta_0) = \mathcal{N}(a_0, Q_0) \end{aligned}$$

Two unknown quantities: the unobservable states θ_t and the hyperparameter $\lambda = (a_0, Q_0, Q, F)$.

The conditional distribution of $\Theta = (\theta_0, \ldots, \theta_T)'$ given the observations $Y = (y_1, \ldots, y_T)'$

$$p(\Theta|Y) \propto \prod_{t=1}^{T} p(y_t|\theta_t) \prod_{t=1}^{T} p(\theta_t|\theta_{t-1})p(\theta_0).$$

Because of the complicated form of the conditional distribution, inference requires some approximation

(Generalized extended Kalman filter and smoother (Fahrmeir, 1992) + GCV).

A possible model for the Alabama data

$$y_t|\phi_t \sim Po(e^{x_t^T\alpha+\phi_t})$$

with:

$$\log(\phi_t) = \omega_t + a_t\text{Tmax} + b_t\text{Tmin} + c_t\text{Hum} + d_t\text{PM}_{10}$$

$$e^{x_t^T\alpha} = 1$$

$\Rightarrow$ leave to the model the decision about which of the covariates show a stochastic trend

By running the procedures we found

- Tmax not significant;
- Tmin ordinary fixed effect;
- Humidity and PM10 time-varying.

Results

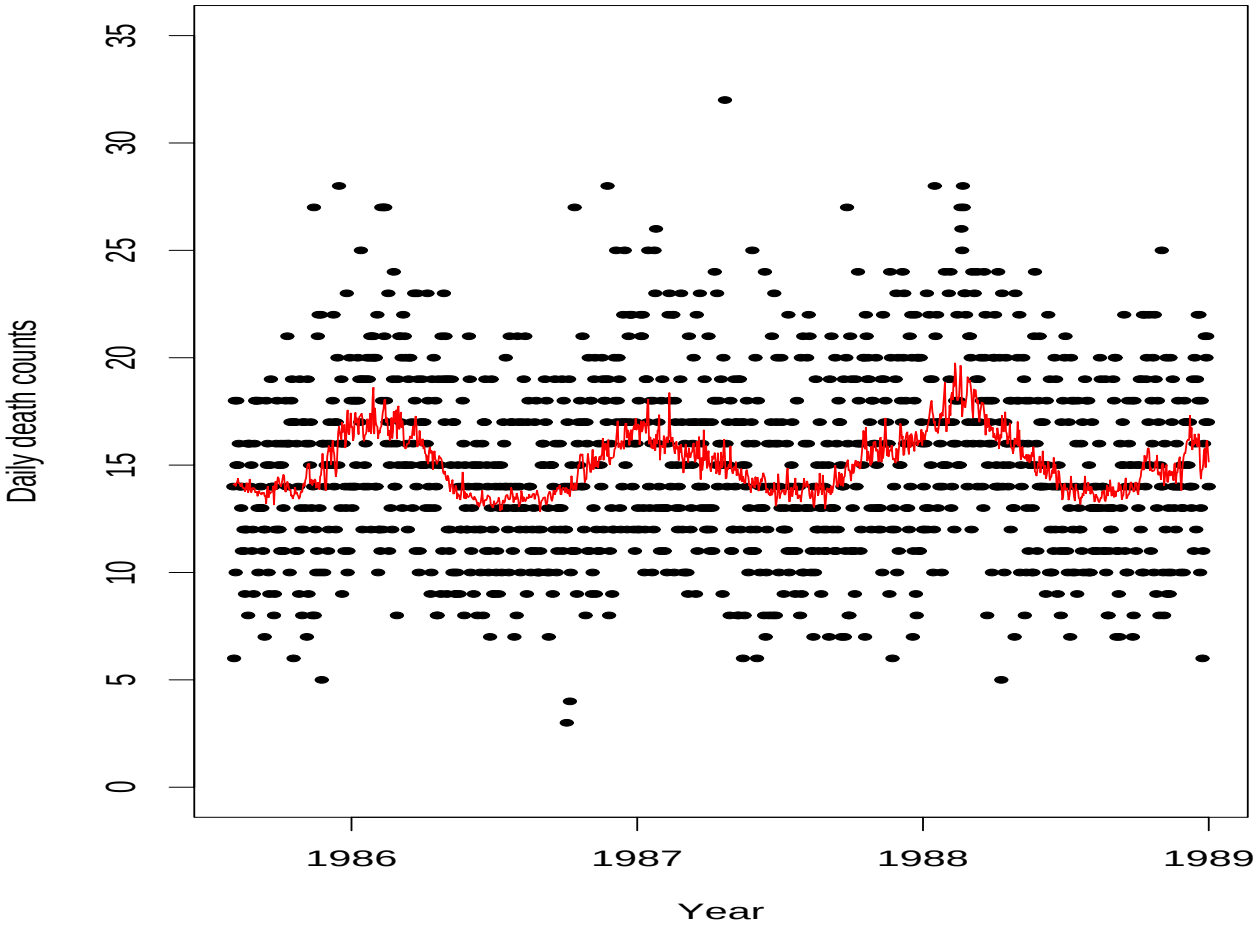

Figure 2: Time series of the observed counts along with fitted values.

Trend

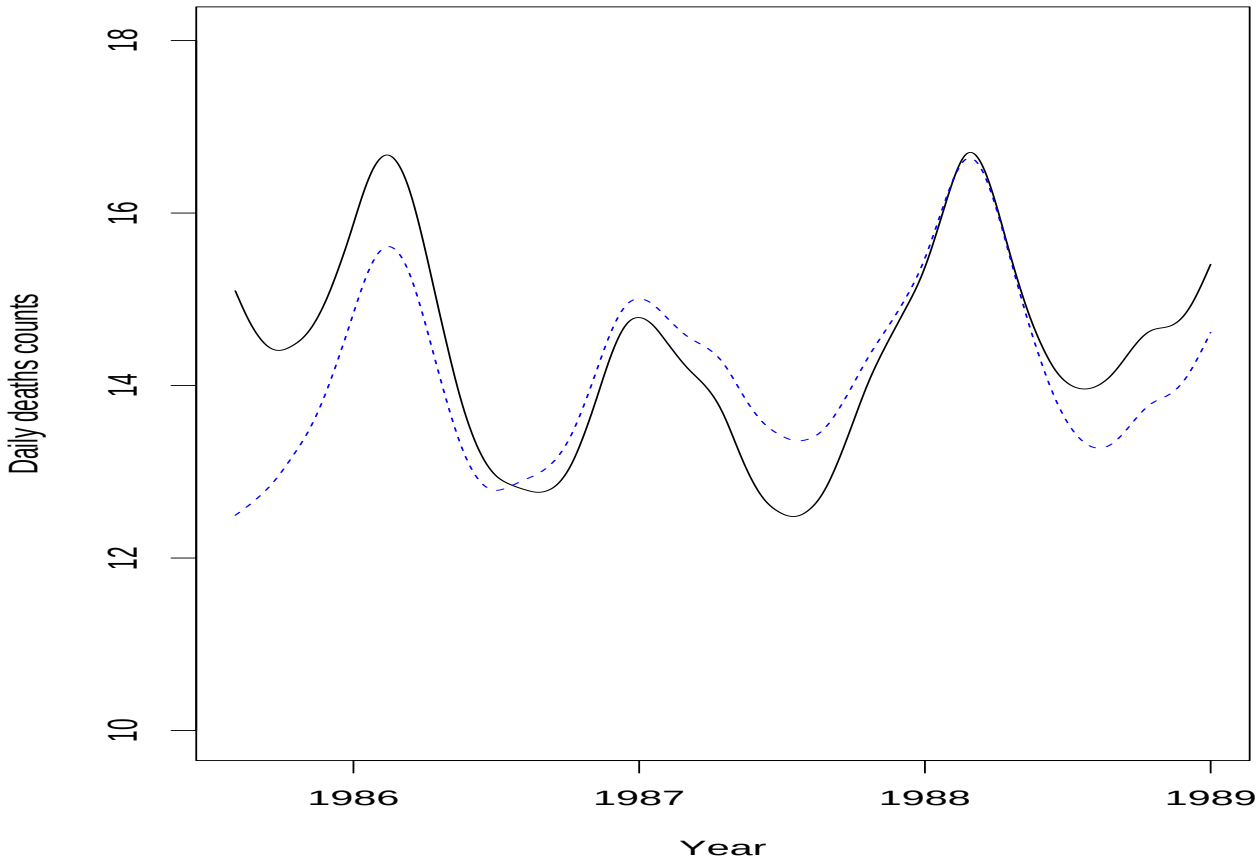

Figure 3: Estimated trend for the DGLM (solid line) and for the semiparametric model: $\log(\mu_t) = \alpha_1 + s(t,k) + \alpha_2 \text{Tmin}_t + \alpha_3 \text{Hum}_t + \alpha_4 \text{PM10}_{t-1}$ (dashed line).

Time varying coefficients

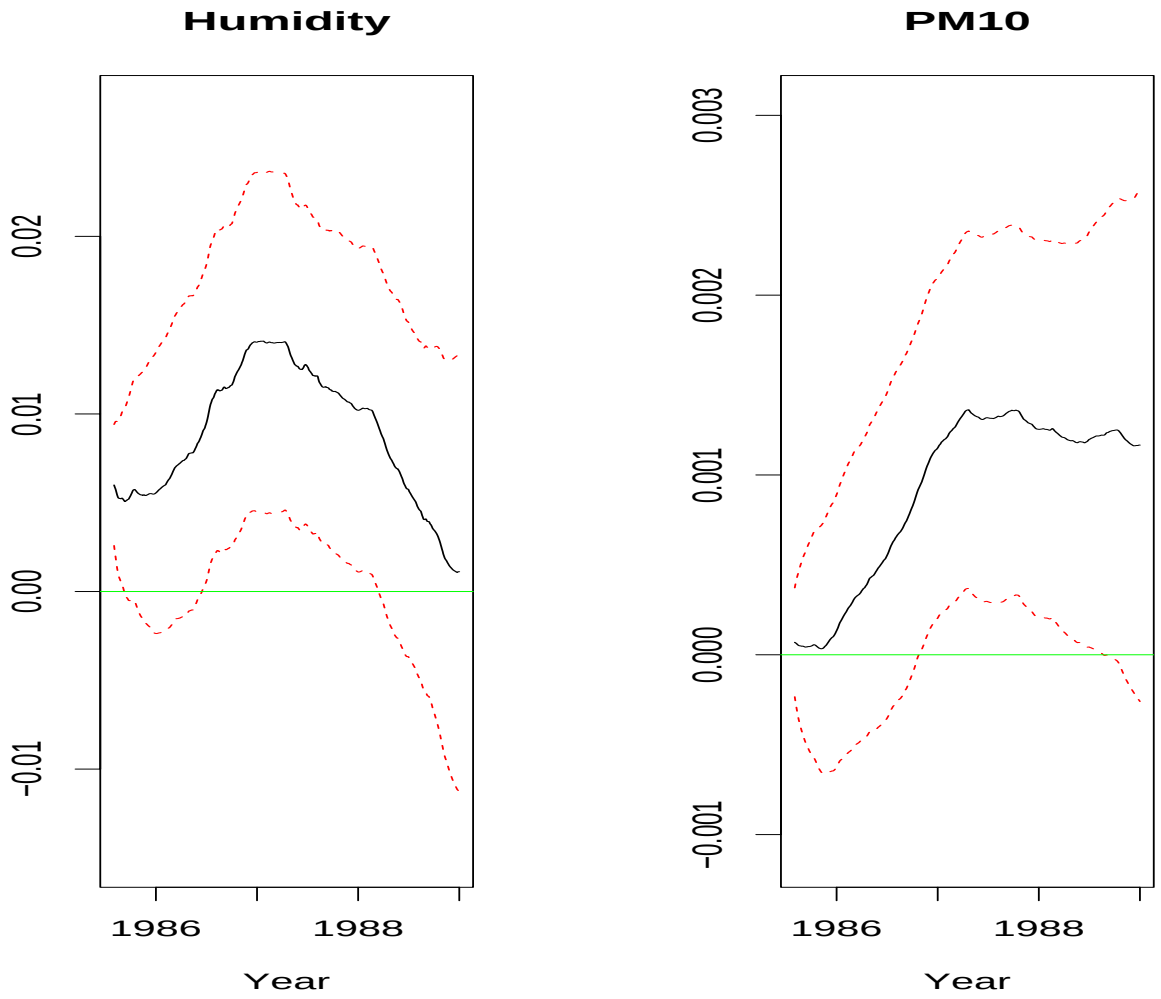

Figure 4: Estimated trajectories for the coefficients of humidity and PM10 with pointwise 95% confidence intervals.

Randomized residuals

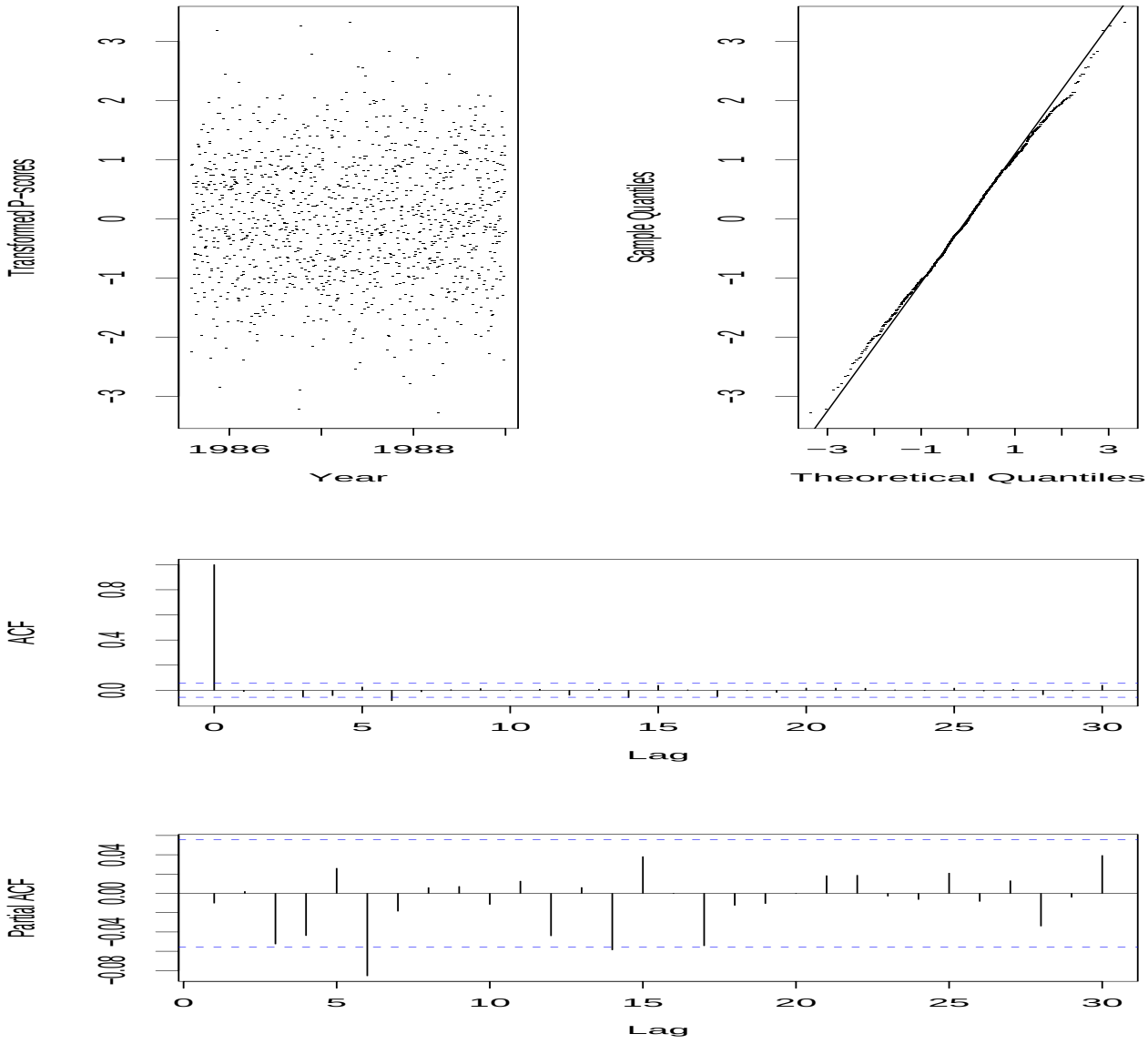

Figure 5: Diagnostic plots on the randomized residuals.

Results and Conclusions

Goods

- Unifying framework for dealing with fixed effects, time varying effects and stochastic trends.
- Definition of lagged effects of covariates is left to the model.
- Possibility to check for possible variations in time of the pollutant's effect.
- Starting point for MCMC inference.

Bads

- No standard software available for fitting the models (R functions + C code).
- Careful choice of initial values for hyperparameters.

Essential References

- Chiogna, M., Gaetan, C. (2002). "Dynamic generalized linear models in environmental epidemiology". *Applied Statistics*. To appear.
- Smith, R.L., Davis, J.M., Sacks, J., Speckman, P., and Styer, P. (2000). "Regression models for air pollution and daily mortality: analysis of data from Birmingham, Alabama". *Environmetrics*, **11**, 719-743.
- 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.
- Fahrmeir, L., Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York.