

Transfer Function Analysis of Epidemiological Data

Un'Analisi di Dati Epidemiologici mediante Modelli a Funzione di Trasferimento

Monica Chiogna

Dipartimento di Scienze Statistiche
Università di Padova
monica@stat.unipd.it

Carlo Gaetan¹

Dipartimento di Scienze Statistiche
Università di Padova
gaetan@stat.unipd.it

Riassunto: Nel presente lavoro si propone l'uso di modelli a funzioni di trasferimento per l'analisi di serie storiche epidemiologiche. La metodologia proposta prevede due fasi: nella prima fase, le serie storiche relative alla risposta ed alle variabili esplicative vengono individualmente depurate dell'effetto di confondimento temporale e meteorologico. Successivamente, si utilizzano le serie dei residui derivanti dalla prima fase per costruire un nuovo modello. In entrambe le fasi, si utilizzano modelli a funzioni di trasferimento selezionati tramite algoritmi genetici. L'utilità della strategia viene esplorata nel contesto degli studi di associazione inquinamento atmosferico–effetti sulla salute.

Keywords: Epidemiological Time Series, Genetic Algorithms, Model Selection.

1. Introduction and background

Time series studies are specially suitable in epidemiology for evaluating short-term effects of time-varying exposures to air pollution on human health. Evidence in the literature shows that findings are fairly robust to alternative choices of the modelling strategy. Nevertheless, various issues raise concern and it is generally agreed that most of the existing studies might share several methodological weaknesses (see, for example, the special issue of *Environmetrics*, vol. 11, issue 6, 2000, and references therein for discussions about the impact on the conclusions of the methodology used in these studies).

In this work, we propose to make use of transfer function models to model association between air pollution and health. Transfer function models are a well-known and useful modelling strategy, but to our knowledge they have not been used in environmental epidemiology. They provide a convenient framework to deal with dynamic features, serial correlation and lagged effects. As the space of models explored by the proposed methodology is very big, and because we do not want to a priori eliminate any potential variables, we propose to make use of Genetic Algorithms to stochastically search the space of models.

We illustrate these ideas by re-analysing data on the relationship between daily non accidental deaths and air pollution in the city of Philadelphia, PA. Section 2 will illustrate the idea behind our modelling approach, whereas transfer function models and Genetic Algorithms for model identification will be illustrated in Section 3. The last two sections of the paper cover the application and summarize our findings.

¹This work was supported by MIUR (Italy) under grant n. MM3208412 : *Statistics in Environmental Risk Assessment*. We wish to thank Guido Masarotto for his useful remarks and for providing us with the computational framework.

2. The modelling rationale

Our goal is to model the relationship between a response time series, and some explanatory time series of interest (the pollutants' concentrations) in the following referred to as determinants, taking into account the intervention of confounders that might alter the measure of association between response and determinants. In epidemiological time series studies, confounding effects are mostly related to temporal trends and meteorological effects. Control of confounding is obtained via regression strategies by inserting the confounding factors into the predictor before considering the determinants of interest. This works well in adequately capturing the underlying seasonal pattern in the response but it does not accommodate the seasonal patterns in each of the determinants.

Therefore, rather than build a model directly between the response time series and the explanatory time series of interest controlling confounding via regression, we propose to build a model between the response and the explanatory time series individually adjusted for temporal and meteorological confounding. Thus at the first stage, to the response and to each of the determinants appropriate transfer function models are fitted having the confounders as input variables. The residual series from these fits are then cross correlated by means of a new transfer function model. By recombining the models for the original variables with the identified model connecting the residuals, a distributed lag models relating the response to the determinants is identified, which may be fitted and checked using standard methodology.

3. The modelling framework

Assume that the output $\{Y(t)\}$ and the inputs $\{X_i(t)\}$, $i = 1, \dots, k$, are stationary processes which are related through the the transfer function model

$$Y(t) = \sum_{i=1}^k \frac{\omega_i(B)}{\delta_i(B)} X_i(t) + \frac{\theta(B)}{\phi(B)} e(t), \quad (1)$$

where $\omega_i(B)$, $\delta_i(B)$, $\theta(B)$, $\phi(B)$, are polynomials in the backward shift operator, B . The noise process $\{e(t)\}$ is a zero-mean stationary process independent of the inputs. We suppose also that the roots of the polynomials $\delta_i(z)$, $i = 1, \dots, k$ and of $\phi(z)$ lie outside the unit circle. We identify a transfer function model $\mathcal{M}$ with coefficients of the polynomials $\theta_{\mathcal{M}}$. As selection criterion of $\mathcal{M}$, we choose the Bayesian Information Criterion (BIC) $V(\mathcal{M}) = \log \left(\sum_{t=1}^T e(t; \theta_{\mathcal{M}})^2 / T \right) + \dim(\mathcal{M}) \log T / T$, where $e(t; \theta) = Y(t) - \hat{Y}(t; \theta)$, $\hat{Y}(t; \theta)$ is the one-step prediction ahead of $Y(t)$, $\dim(\mathcal{M})$ is the number of unknown coefficients in (1) and $\hat{\theta}_{\mathcal{M}} = \operatorname{argmin}_{\theta} \sum_{t=1}^T e(t; \theta)^2$ is the prediction error estimate.

As well known, $\hat{\theta}_{\mathcal{M}}$ cannot be computed analytically. To solve this nonlinear least-squares problem we use a Gauss-Newton type algorithm

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \lambda_n \left[\sum_{t=1}^T J(t; \hat{\theta}_n) J(t; \hat{\theta}_n)^T \right]^{-1} \sum_{t=1}^T J(t; \hat{\theta}_n) e(t; \hat{\theta}_n)$$

where $0 < \lambda_n \leq 1$ and $J(t; \hat{\theta}_n)$ is the Jacobian vector $\partial \hat{Y}(t; \theta) / \partial \theta$. Note that $\hat{\theta}_{n+1}$ is the least square solution of $J(t; \hat{\theta}_n)^T \hat{\theta}_n - \lambda_n e(t; \hat{\theta}_n) = J(t; \hat{\theta}_n)^T \theta$, $t = 1, \dots, T$.

This remark allows us to couple the estimation step and the selection one. More precisely, we propose this identification procedure: (1) choose a starting value $\hat{\theta}_0$; (2) solve the least square problem and get θ_{n+1}^* ; (3) apply a backward stepwise selection procedure to θ_{n+1}^* according to the selection criterion $V(\mathcal{M})$ and obtain $\hat{\theta}_{n+1}$; (4) set $n = n + 1$ and return to 2 until a converge criterion is met.

This procedure speeds up the identification task but it depends on the starting point $\hat{\theta}_0$, so it can fail to find the optimum of $V(\mathcal{M})$. We propose to overcome this limitation using Genetic Algorithms (GAs) in the same spirit of Gaetan (2000). Our implementation of GAs is based on incremental reproduction and a ranking-based selection scheme (see Masarotto (2001) for details).

4. Modelling in practice

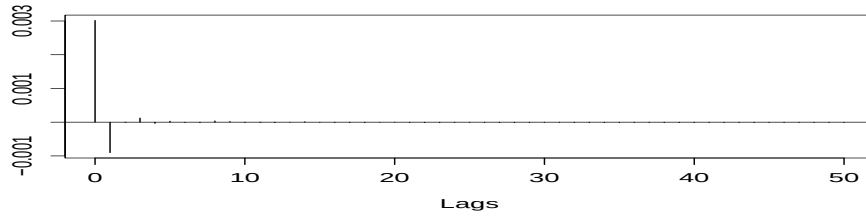
In the present analysis, we have used the same data set as in Kim et al. (1999). The data set consists of 14 years (1974–1988) of daily deaths counts with associated measurements of temperature and dewpoint and of five pollutants: total suspended particulate (TSP), sulphur dioxide (SO₂), nitrogen dioxide (NO₂), carbon monoxide (CO) and ozone (O₃) (refer to Kim et al. (1999) for summary statistics for the variables). Due the nature of the response, a Poisson regression model seems a natural way to explain the variability, but, as the mean number of counts is sufficiently high, we can consider the square root transformation of the response, in the following denoted as $Y(t)$, and move to a linear model. This allows us to connect to the transfer function methodology.

The autocorrelation functions of the square root of daily counts and of the pollutants highlight some non-stationarity in the mean of the series and different cyclical patterns. Differentiating the series leads to more stable results, although differences in the periodic behaviours still persist. This agrees with our intuition that the series need different seasonal adjustment.

Therefore, in the first phase of the model construction, we fitted a transfer function model to the square root of daily counts and to each pollutant with inputs temperature (T_1 =temperature below 75° F, T_2 =temperature above 75° F), dewpoint (D), with all variables differentiated, using the BIC criterion. To explore the space of possible models, we used both standard stepwise selection procedures and GAs. Results from this first phase confirmed that temporal confounding was different on the response and on the pollutants. There was evidence of short term (about 7 days) seasonality in all pollutants but SO₂ and O₃; the same effect was not so evident in the square root of the deaths. Effect of meteorology always had a strikingly significant effect on the output series, although this was different for response and for pollutants. This confirmed that meteorology accounts for most of the seasonality and supported the idea of separate adjustments.

After meteorological adjustment, we moved on to the second phase of the model construction, i.e. to the fitting of a transfer function model to the estimated residuals from Phase 1. In this second stage, only SO₂ showed a significant effect. The current day value for SO₂ resulted to be the most predictive, which agrees with Kelsall et al. (1997). This is also quite consistent with findings of Kim et al. (1999). Therefore the final model, using GAs, looked as follows: $\hat{\varepsilon}(t) = 0.003\hat{\varepsilon}'(t)$, where $\hat{\varepsilon}(t) = \alpha(B)[Y(t) - \beta_2(B)T_2(t) - \beta_3(B)D(t)]$ with $\alpha(B) = (1-B)(1-0.041B^3-0.374B^5-0.341B^6)/(1-0.954B-0.362B^5+0.318B^7)$, $\beta_2(B) = 0.032/(1-0.587B)$ and $\beta_3(B) =$

Figure 1: Impulse responses to a unit pulse input of sulphur dioxide.



$(0.005 + 0.007B + 0.003B^2 - 0.003B^3)/(1 - 0.856B)$ and $\varepsilon'(t) = \alpha'(B)[SO_2(t) - \beta'_1(B)T_1(t) - \beta'_2(B)T_2(t) - \beta'_3(B)D(t)]$ with $\alpha'(B) = (1-B)(1-0.298B)/(1-0.953B)$, $\beta'_1(B) = (0.239 + 0.329B^2)/(1 - 0.511B^2)$, $\beta'_2(B) = 1.078$ and $\beta'_3(B) = (0.119 + 0.323B)/(1 + 0.285B)$. By recombining the models for the original variables with the identified model connecting the residuals, the obtained impulse response function to a unit pulse input of sulphur dioxide is shown in Fig. 1. Inspection of the figure shows the positive effect of SO_2 at lag 0 on the square root of counts. On the other hand it shows a negative association between pollutant and counts at lag 1. This result might be consistent with the mortality displacement hypothesis.

5. Discussion

In this work we discuss the use of transfer function models in environmental epidemiology and discuss a computational method for selecting a transfer function model based on Genetic Algorithms. We have found transfer function models a very flexible tool to deal with many variables, lagged effects, serial correlation. Genetic search almost always picked out models different from those resulting from standard stepwise selection procedures and preferable in terms of BIC criterion. The use of the two-phase modelling strategy allowed us to achieve a better control of the multicollinearity problem. Moreover, separate adjustments for temporal and meteorological confounding enhanced understanding of possible sources of multicollinearity among pollutants. Results appeared consistent with finding of multi-pollutant studies, marking SO_2 as the most relevant pollutant in the association with mortality.

References

- Gaetan, C. (2000). Subset ARMA model identification using genetic algorithms, *Journal of Time Series Analysis*, 21, 559–570.
- Kelsall, J., Samet, J., Zeger, S. and Xu, J. (1997). Air pollution and mortality in Philadelphia, 1974–1988, *American Journal of Epidemiology*, 146, 750–762.
- Kim, Y., Spitzner, D., Zhang, Z., Smith, R. and Fuentes, M. (1999). Accounting for multiple pollutants in pollution-mortality studies, *ASA Proceedings of the Biometrics Section*, 1–10.
- Masarotto, G. (2001). Automatic identification of ARMA, ARMAX and Transfer function models, Working Paper of Department of Statistics, University of Padua.