Web Working Papers
by

The Italian Group of Environmental Statistics

Gruppo di Ricerca per le Applicazione della Statistica
ai Problemi Ambientali

www.graspa.org

Time series studies of air pollution and mortality: is semi-automated model selection possible?

Monica Chiogna e Carlo Gaetan

GRASPA Working paper n.17, Settembre 2003

Working paper GRASPA (2003), **17**, *pp.* 1–13

Time series studies of air pollution and mortality: is semi-automated model selection possible?

By MONICA CHIOGNA

Dipartimento di Scienze Statistiche, Università di Padova,
Via Cesare Battisti, 241, I-35121 Padova, (Italy).
e-mail: monica@stat.unipd.it

and CARLO GAETAN

Dipartimento di Scienze Statistiche, Università di Padova,
Via Cesare Battisti, 241, I-35121 Padova, (Italy).
e-mail: gaetan@stat.unipd.it

Abstract

Time series regression models are specially suitable in epidemiology for evaluating short-term effects of time-varying exposures. The objectives of this paper are twofold: 1) to apply transfer function models for regression analysis of epidemiological time series; 2) to explore the potential of semi-automated or automated approaches for model construction. The ideas are illustrated by analysing data on the relationship between daily non accidental deaths and air pollution in the 20 US largest cities.

Some key words: Environmental epidemiology; Transfer Function Models; Genetic Algorithms

1. Introduction and background

The literature suggests that a range of health effects occurs at levels of pollutants below US and European regulatory standards. All analyses of the short-term relationship between atmospheric pollution and human health outcomes rely on regression analysis in which, for one city, time series of health outcomes (daily mortality or hospital admissions counts), are regressed over a number of exposure factors (such as air pollution concentrations) and of confounding factors (such as meteorology) measured for the same time period.

Despite the simplicity of the approach, there are many variations in the precise methodology adopted. The models that have been used differ in the parametric assumptions, in their approaches to estimating natural temporal trends, in the detail with which confounding effects are treated, in the extent to which displacement of the events is taken into account.

Therefore, two main critics appear frequently in the literature. The first one is related to a percieved structural methodological weaknesses, which, ultimately, has to do with an inadequate statistical control of the dependence structure of the data. The second critic concerns the many *ad hoc* choices that the construction of such models requires. It is often the case that crucial choices on the structure of the regressor (such as the degree of smoothing of non parametric components) exploit what is reasonably expected about the way in which the data are generated. These choices turn out to be quantitative judgements placed in the model which can only be

checked by re-running analyses under slightly different scenarios. Unfortunately, results from this procedure, if performed, are hardly ever reported in the literature.

On the other side, data-driven model selection does not seem to be adequate in this context, and different model selection strategies may lead to different models and conclusions. An attempt to overcome this limitation has been performed by Clyde (2000), which instead of selecting one best model, adopts Bayesian model averaging (Hoeting *et al* (1999)) for combining predictions and inferences from a set of competing models.

The effort required for performing model selection is amplified by the fact that is becoming current practice pooling evidence from different studies. Recent meta-analysis studies (Dominici *et al* (2000), Samet *et al* (2000), Biggeri *et al* (2001), Atkinson *et al* (2001)) showed that combination of data from disparate sources provides additional statistical power to the analysis, that it is not available in single site analyses. Clearly, selection of a model to be used in the meta analysis becomes rapidly more complicated as the number of cities increases.

In this paper, we wish to investigate the potential of transfer function analysis in simultaneously providing a convenient modelling approach for dealing with the complex dependence structure shown by these data and an affordable computational framework to allow a data-driven model selection. By performing a (semi-)automated model selection we want to limit the number of *ad hoc* choices and leave the data to highlight the most relevant features.

Transfer function models are a well-known and useful modelling strategy, but, to our knowledge, they have never been used in environmental epidemiology. They are able to deal with many potential covariates, lagged values, seasonal interactions, dynamic features, appearing therefore a promising vehicle to approach this modelling task. Automated model selection, however, is not straightforward because enumeration over all possible models is not feasible. To tackle the data-driven model selection, we propose to make use of Genetic Algorithms (Holland (1975)) to stochastically search the space of models.

We illustrate these ideas by analysing data on the relationship between daily non accidental deaths and air pollution in the 20 largest US cities (Dominici *et al* (2000), Samet *et al* (2000), Samet *et al* (2000)). With respect to the application, our final objective is not to produce a measure of increase in risk by an increase in the pollutant value, as it is often the case in these studies. From an applied point of view, our objective is to evaluate the extent to which this class of models, coupled with a semi-automated model selection procedure, offers information about how to build a sensible "common" model to be used in combining evidence from different sources.

The main features of the data are illustrated Section 2. Section 3 briefly recalls transfer function models. Section 4 shows how Genetic Algorithms can be used to identify transfer function models. The last two sections of the paper cover the application and summarize our findings.

2. The data

The data that we consider come from the National Morbidity, Mortality and Air Pollution Study (NMMAPS, Dominici *et al* (2000), Samet *et al* (2000), Samet *et al* (2000)), to which we refer for further details about sources of the data. NMMAPS reports a nationwide study performed in the United States of America of acute health effects of air pollution on morbidity and mortality. Data are public and available at the URL `http://ihapss.biostat.jhsph.edu/data/`.

In what follows, we will analyse the association between daily changes in the concentration of carbon monoxide (CO) and daily number of deaths in the 20 US largest cities. Some summary statistics about the data are reported in Table 1.

Table 1. *Summary statistics for the 20 largest cities (1987–1984).*

Location	Population*	$\bar{y}$	$\bar{x}_{CO}$ (ppb)	CO_{miss} (%)	$\bar{x}_{temp}$ (°F)	$\bar{x}_{dew}$ (°F)
Los Angeles	8863164	148	1057	0.00	68	59
New York	7510646	191	2043	0.07	55	42
Chicago	5105067	114	790	0.00	50	40
Dallas-Fortworth	3312553	49	741	0.00	66	52
Houston	2818199	40	886	0.00	69	59
San Diego	2498016	42	1102	0.00	63	53
Santa Ana-Ana	2410556	32	1243	0.00	63	49
Phoenix	2122101	38	1259	0.10	75	41
Detroit	2111687	47	663	0.00	50	40
Miami	1937094	44	1061	0.10	77	67
Philadelphia	1585577	42	1179	0.10	56	44
Minneapolis	1518196	26	1181	0.14	46	35
Seattle	1507319	26	1779	0.00	52	44
San Jose	1497577	20	1041	0.00	60	45
Cleveland	1412141	36	855	0.85	51	41
San Bernardino	1412140	20	1032	0.00	67	35
Pittsburg	1336449	38	1220	0.07	52	41
Oakland	1279182	22	914	0.00	58	50
San Antonio	1185394	20	1011	1.06	69	56
Riverside	1170413	20	1117	0.03	67	35

*County Population, as derived from NMMAPS, $\bar{y}$= average daily deaths, $\bar{x}_{CO}$: average of daily CO levels, CO_{miss}: percentage of days with missing CO values, $\bar{x}_{temp}$: average of daily mean temperatures, $\bar{x}_{dew}$: average of daily dewpoint temperatures.

The reason for analysing CO instead of PM_{10}, which is the compound of central interest in NMMAPS, is related to the number of missing values. As shown in Table 1, missing values are almost absent in the dataset, and this is an essential feature for applying transfer function models. When missing, data have been estimated by mean of a nonparametric smoother.

Distribution of the original daily counts and average levels of CO are shown in Fig 1 and Fig 2. As far as daily mortality is concerned, the picture highlights a great deal of variability amongs cities, with New York city showing the highest daily mortality. Two other cities (Los Angeles and Chicago) show distributions with remarkably high mean daily deaths counts. Obviously, this reflects also the ranking of the cities in terms of County population (note that for Los Angeles and Chicago it was possible to separate the city and the county deaths, so that data reported for daily mortality reflects only city daily deaths - see Samet *et al* (2000), p. 9). The same ranking is not visible in the distribution of the pollutant, as smaller cities show levels of pollution

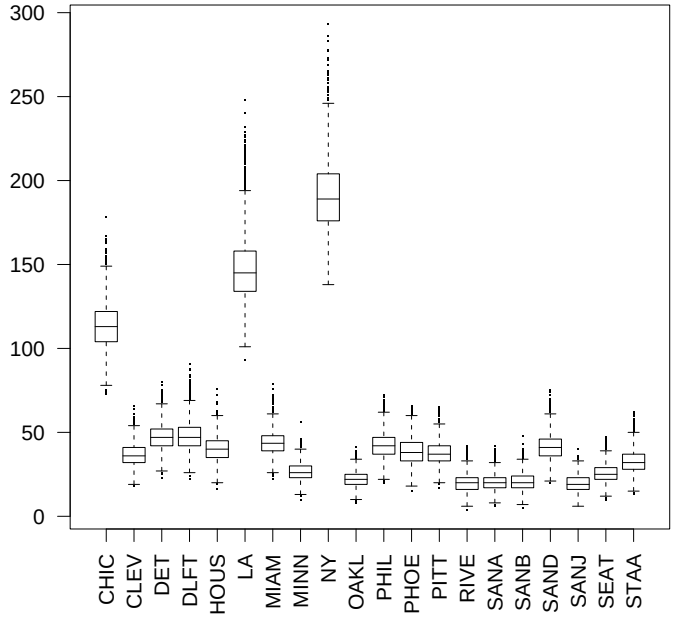

Figure 1. Distribution of daily mortality counts for the 20 cities, 1987–1994.

competing with levels recorded in bigger cities (compare for example Seattle with Chicago).

As far as the effect of the pollutant is concerned, NMMAPS reports positive significant effects of CO at the usual lags (0,1,2).

3. Transfer Function models

In the transfer function setting, the output $\{Y(t)\}$ results from the contribution of some inputs $\{X_i(t)\}$, $i = 1, \ldots, k$, which are stationary processes, and of a noise process $\{e(t)\}$, which is a zero-mean stationary process independent of the inputs. All the ingredients are related via the Transfer Function (TF) model

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

where $\omega_i(B) = \omega_{i0} - \omega_{i1}B - \ldots - \omega_{ir_i}B^{r_i}$, $\delta_i(B) = 1 - \delta_{i1} - \ldots - \delta_{is_i}$, $\theta_i(B) = 1 - \theta_{i1}B - \ldots - \theta_{iq}B^q$, $\phi_i(B) = 1 - \phi_{i1}B - \ldots - \phi_{ip}B^p$, are polynomials in the backward shift operator, B (i.e. $BY(t) = Y(t-1)$), with degrees r_i, u_i, q, p respectively. Moreover, the roots of the polynomials $\delta_i(z)$, $i = 1, \ldots, k$ and $\phi(z)$ are supposed to be outside of the unit circle.

A TF model $\mathcal{M}$ is identified by specifying the degrees and the coefficients of the polynomials. The identification consists of two steps: 1) selection of the degrees of the polynomials; 2)

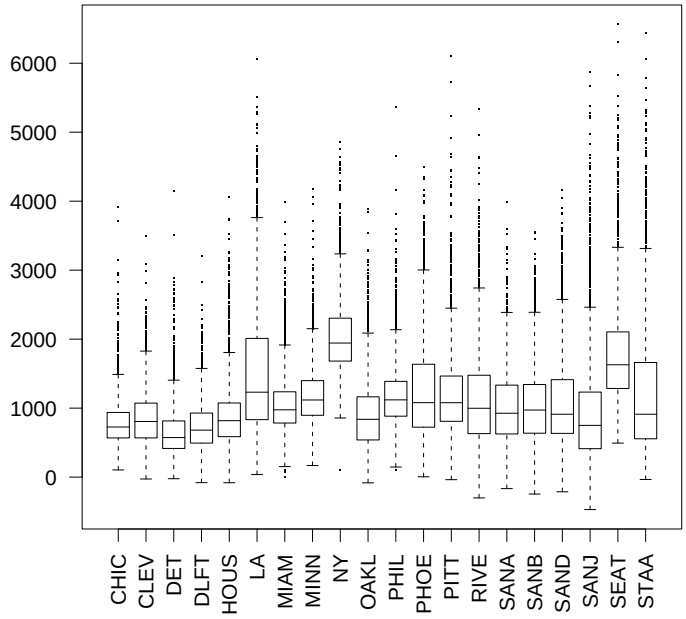

Figure 2. Distribution of daily concentration of CO (ppb) for the 20 cities, 1987–1994.

estimation of the vector of the unknown coefficients $\theta_{\mathcal{M}}$.

The widely used selection procedure for the degrees (Box *et al* (1994)) requires examining cross-correlation functions between (prewhitened) inputs and output. This step involves rather subjective choices about the degrees of the polynomials, especially r_i and u_i.

Usually, the vector $\theta_{\mathcal{M}}$ is estimated by minimising the conditional sum of squares of the prediction errors

$$s(\theta) = \frac{\sum_{t=t^*+1}^{T} e(t;\theta)^2}{T^*}, \tag{2}$$

where $t^* = p + \max_{i=1,k}\{\sum_{j=1}^{k} u_i, r_i + \sum_{j\neq i}^{k} u_i\}$, $T^* = T - t^*$, $e(t;\theta) = Y(t) - \hat{Y}(t;\theta)$ and $\hat{Y}(t;\theta)$ is the one-step prediction error of $Y(t)$.

The selection of $\mathcal{M}$ is usually based upon the Kullback-Leibler information criterion. In this case, it can be approximated by the penalized function

$$V(\mathcal{M}) = \log s(\hat{\theta}_{\mathcal{M}}) + m\frac{c(T^*)}{T^*}, \tag{3}$$

where $m = \dim(\hat{\theta}_{\mathcal{M}})$ is the number of unknown coefficients in (1), $\hat{\theta}_{\mathcal{M}}$ is the prediction error estimate and $c(T^*)$ is a function of the number T of observations. This formulation encompasses many criteria, namely the Akaike information criterion (Akaike (1973)) for $c(T) = 2$, the Bayesian information criterion (Schwartz (1978)) for $c(T) = \log T$ and the Hannan-Quinn criterion (Hannan and Quinn (1979)) for $c(T) = 2\log\log T$.

4. Transfer function model identification by means of a genetic algorithm

As an automated model selection procedure has to make a choice from amongst any set of competing models, the number of possible candidates will get huge as soon as complexity of the model moderately increases. Therefore, automatic identification of a model poses serious computational difficulties. We propose to tackle this task by making use of genetic algorithms (GAs), in the same spirit of Gaetan (2000). GAs are a family of computational strategies inspired by evolution Holland (1975). These algorithms encode a potential solution to a specific problem on a simple chromosome-like data structure and apply recombination operators to a population of these structures so as to preserve critical information. GAs are often viewed as function optimizers, although the range of problems to which genetic algorithms have been applied is quite broad. Our implementation of the GA is based on incremental reproduction and a ranking-based selection scheme (Whitley (1989)). Let q^*, p^*, r_i^*, u_i^* $i = 1, \ldots, k$ the maximum degree allowed for each polynomial and $M = p^* + q^* + \sum_{i=1}^{k}(r_i^* + u_i^*)$. We identify a model $\mathcal{M}$ with a chromosome (string) of M genes. The value of a gene is a component of $\theta_{\mathcal{M}}$. Note that this setting includes consideration of TF subset models, i.e. models in which polinomials representing the lag structure can be incomplete. This means that cardinality of the model space is 2^M.

Our GA for transfer model identification is the following (see Chiogna *et al* (2003) for more details).

1. Create a random population of N models.
2. Compute for all models the fitness, namely the criterion (3). Assign to each model its rank value.
3. Select two "parent" models according to their rank to form an "offspring". The "offspring" $\mathcal{M}^{(O)}$ of two parents $\mathcal{M}^{(P_1)}$, $\mathcal{M}^{(P_2)}$ is generated according to
$$\theta_j^{(O)} = \frac{\theta_j^{(P_1)} + \theta_j^{(P_2)}}{2}, \quad j = 1, \ldots, M\cdot$$
4. Apply to the "offspring" the mutation operator. This operator chooses a gene with probability p_m and then generates a value from normal distribution with mean 0 and variance σ^2. For the coefficients in $\omega_i(z)$ the variance is set to $\sigma^2 = s_Y^2/s_{X_i}^2$; for the others polynomial we have chosen $\sigma^2 = 0{\cdot}1$. In principle, the roots of $\phi(z)$ and $\delta_i(z)$ in the "offspring" do not lie outside the unit circle. So polynomials are changed in order to satisfy this constraint.
5. The coefficient values found in the previous steps are the starting value in the numerical optimisation of (2). After the evaluation of the offspring's fitness, if the estimated model is different from the population's ones, this replaces the least fit individual according to criterion (3). Otherwise a new model is generated and its coefficients are generated as in the mutation stage.
6. Return to 3 until a convergence criterion is met.

The computational burden lies on optimization of (2) because $\hat{\theta}_{\mathcal{M}}$ cannot be computed analytically. We use a Gauss-Newton type algorithm to solve this nonlinear least-squares problem, namely

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

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

$$J(t;\hat{\theta}_n)^T\hat{\theta}_n - \lambda_n e(t;\hat{\theta}_n) = J(t;\hat{\theta}_n)^T\theta, \quad t = t^*,\ldots,T\cdot \tag{4}$$

This remark allows us to couple the estimation step and the selection one in order to speed up further the identification procedure. More precisely, at step 5 we solve the least square problem (4) and get θ^*_{n+1} and we apply a statistical selection procedure to θ^*_{n+1} according to the selection criterion (3) and obtain $\hat{\theta}_{n+1}$. Various statistical selection procedures (Draper and Smith (1998)) can be chosen. For example, if M is less than 30 a branch-and-bound technique (Miller (1990)) is feasible, but we have found that an iterative stepwise selection perform just as well and it does not require huge allocation of computer memory (Chiogna *et al* (2003)).

5. Results

In this section, we illustrate the results obtained in our analysis. As described in the introduction, our objectives were twofold: 1) to explore the potential of transfer function analysis for the study of effects of air pollution and health by making use of real data; 2) to explore the potential of (semi-)automated or automated approaches for model construction. Therefore, in what follows we will split considerations about the algorithmic performances of the modelling strategy and considerations about assessment of the relation between mortality and air pollution in the 20 largest cities in two separate Sections.

5.1. Modelling strategy

Denote the daily counts by $\{C(t)\}$, $t = 1,\ldots,T$ and by $\{X_i(t)\}$ the time series for input i. In our example $\{X_1(t)\}$, $\{X_2(t)\}$, $\{X_3(t)\}$, are the pollutant, temperature and dew point time series, respectively. Due the nature of $C(t)$, a Poisson regression model would be a natural way to explain the variability. However, as the mean number of counts is sufficiently high, we can safely consider the transformation $Y(t) = \sqrt{C(t)}$ and move to linear models. This is supported by the findings of Smith *et al* (2000) and allows us to connect to the transfer function methodology.

Figure 3 shows the autocorrelation function of $Y(t), X_1(t)$, and $X_2(t)$, $t = 1,\ldots,T$, for the city of Los Angeles. The plot highlights some non-stationarity in the mean of the series, long persistence and different cyclical patterns. Computing the first differences the series leads to more stable results, with a slight indication of over-differentiation, as shown in Figure 4. Across cities, indications about periodicities of the response and covariates varied, although not drammatically.

To offer our model section procedure a great deal of flexibility, we chose the following model setting:

$$Y(t) = \sum_{i=1}^{3}\omega_i(B)X_i(t) + \frac{\theta(B)}{(1-B)(1-B^7)\phi(B)}e(t),$$

where $r_i = 3$, $i = 1,\ldots,3$ and $p = q = 7$. As model selection criterion we chose BIC. This formulation allows to take into account short term seasonal patterns and long term trends. Moreover, it allows to incorporate lagged values of the inputs. Based on evidence from the literature, we considered that the first three lags were sufficient to catch delayed effects of the covariates. Note that, despite the relative simplicity of this model formulation, cardinality of the model space is around $6{\cdot}7 \times 10^7\cdot$

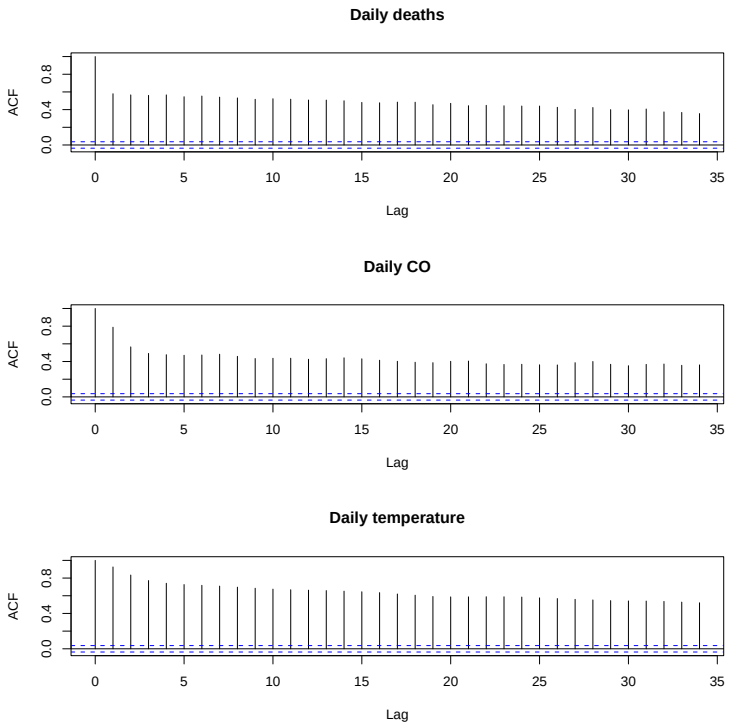

Figure 3. Autocorrelation functions of time series of square root of deaths counts, CO and average daily temperature.

At the end of the model selection, in 8 cities found the search strategy a significant effect of CO. In all the 20 cities, the selection procedure adopted first order differences for the input and the output series. Details on the selection of the best lags for the models in the 8 cities with significant effects are summarized in Table 2. Differencing the data led to a good control of the main temporal patterns. On the differenced data, weekly ciclicity did not seem to be relevant. Only Dallas-Forthworth showed a residual weekly ciclicity, although the Box-Ljung test detected presence of autocorrelation in the residuals (p-value=$1{\cdot}6 \times 10^{-13}$), showing that the model was not able to whitening the residuals, and therefore the temporal control was not fully satisfying.

It is not surprising that on 12 cities the model selection procedure did not spot a significant effect of CO. This group comprises cities in which there is actually no evidence of a significant effect of the pollutant, and cities in which the identified best model very closely competed in terms of goodness of fit (as measured by AIC, BIC, HQ) with a model including the pollutant as significant, as we *a posteriori* checked by re-fitting *ad hoc* competing models. This makes the number of 8 cities comparatively a high number of cities with significant effect and a strong evidence in favour of an association between the pollutant and accidental deaths.

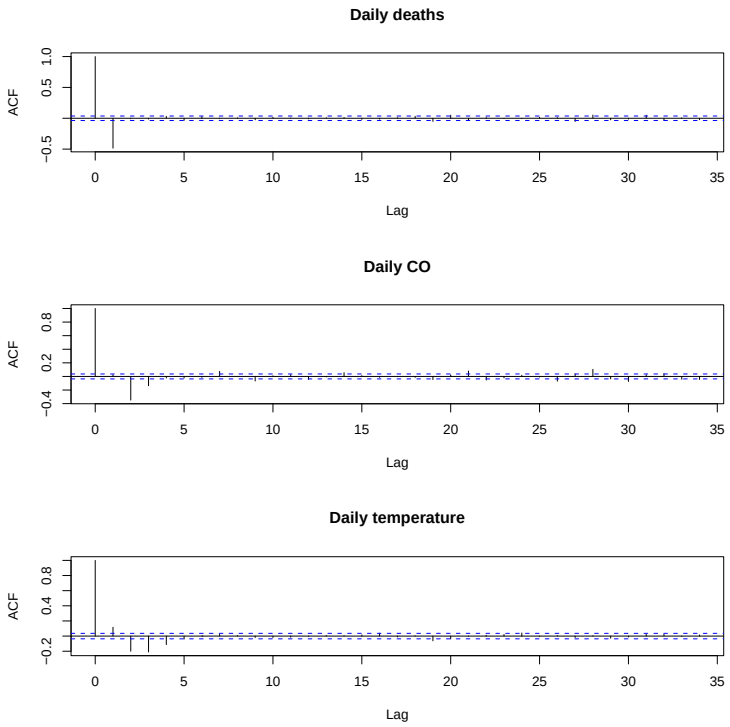

Figure 4. Autocorrelation functions of time series of differenced square root of deaths counts, *CO* and average daily temperature.

Table 2. *Results of the automated model selection procedure for the cities showing a significant effect of CO. Structure of the lags for each explanatory variable. When missing, the variable was not reported as being significant.*

Location	Temperature	Dewpoint	*CO*	Noise
Dallas-Forthworth	0	2	0	1,2,5,7
Los Angeles	1,2	2	1,3	1
Miami	3		3	1
Oakland	3		1	1
Phoenix	3	0,1	2	1
San Diego			1	1
San Jose			1	1
Santa Ana-Ana		2	1	1

5.2. Air pollution and mortality

To perform the meta-analysis task, we adopted the strategy of fitting the same common model to all the cities and to combine evidence resulting from the model fitting. Based on the output

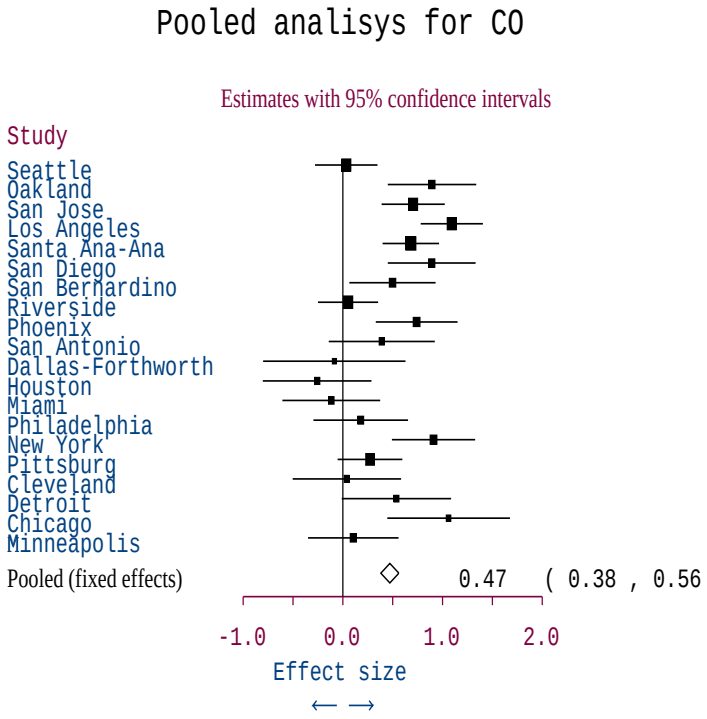

Figure 5. Results of the pooled analysis under the fixed effects model (coefficients are multiplied by 10^4).

from the automated model selection procedure, we decided to fit the following common model:

$$Y(t) = \omega_{11} X_1(t-1) + \omega_{23} X_2(t-3) + \omega_{32} X_3(t-2) + \frac{(1-\theta B)}{(1-B)} e(t).$$

The common model allowed to detect a significant effect in 11 cities, providing further evidence for what we already said at the end of the previous paragraph. In the meta-analysis, the estimates for CO for each city were combined using fixed and random effects models (Normand (1999)). City-specific and pooled estimates are represented graphically in Figure 5 and Figure 6. Beeing the values of the coefficients very small, for graphical convenience all values have been multiplied by 10^4.

As expected, the confidence intervals were widest under the random effects model, and narrowest under the fixed effects model. Nevertheless, differences in point estimates were negligible. A geographical gradient in value for the effect is visible, with Seattle and Minneapolis distinguishing from the remaining cities. Estimates are significant in Southern California and in the Southwest, become not significant moving to the Southeast, and return significant moving to the Northeast and industrial Midwest. This agrees with the effects found for PM_{10} from NMMAPS. These findings confirm the ability of TF models in detecting the association between the output and the inputs and in giving useful indications to construct sensible models for tackling this complex modelling task.

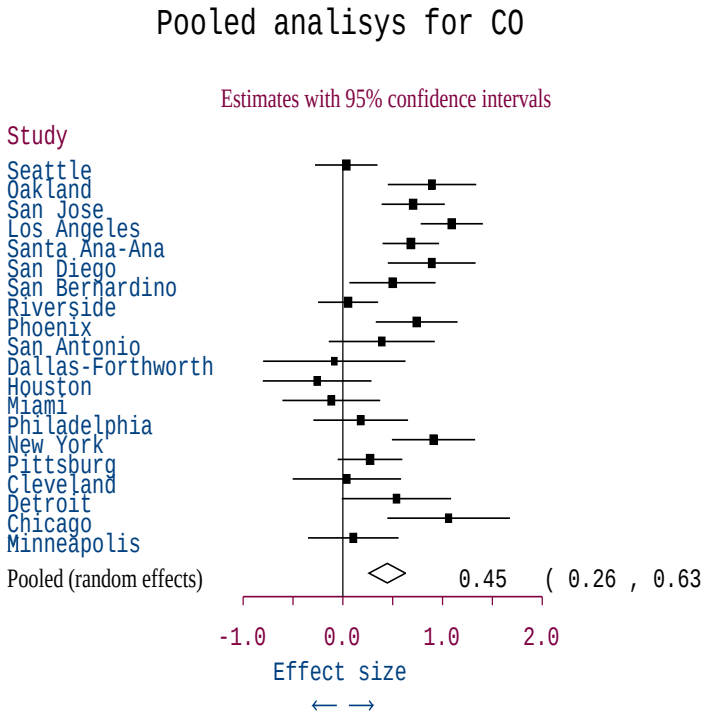

Figure 6. Results of the pooled analysis under the random effects model (coefficients are multiplied by 10^4).

6. Discussion

In this work, we discuss the use of transfer function models in environmental epidemiology, in particular emphasizing their ability to provide realistic information for performing analyses in situations in which evidence is available from a variety of sources. We have found transfer function models a very flexible tool to deal with many variables, lagged effects, serial correlation. Semi-automated model selection, which we have approached via a genetic search over the model space, is challenging and provides useful information for combination of studies from disparate sources.

Clearly, there are some drawbacks related to this paradigm. First of all, this approach amplifies computationally the programming effort that it is typically involved in similar applications. Although this is disadvantageous, advantages include the conceptual appeal of limiting difficult and not always transparent *ad hoc* selections of the structure of the common model. Secondly, the obvious probabilistic model associated with transfer function analysis is a Gaussian model, which implies that the original data need to be transformed. In most applications, which focus on cities with a relevant average number of daily events, transforming the data does not poses any particular problem. Nevertheless, it could cause some difficulties in sites with a small number of daily events.

In spite of these objections, we feel that the transfer function approach offers the modellers realistic information to be used in subsequent analyses.

Acknowledgement

This work was partially supported by MIUR (Italy) grant MM3208412: "Statistics in environmental risk assessment". A preliminary version of this paper was presented at XLI Riunione Scientifica della Societa' Italiana di Statistica, Milan (Italy), 5-7th June 2002 . We wish thank the discussant of the Section for her useful comments. We are also indebited with Guido Masarotto for providing us the computational framework.

References

Akaike, H., (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov, B., Csaki, F. (Eds.), *2nd International Symposium on Information Theory*. Akadémiai Kiadó, Budapest, 267–281.

Atkinson R.W., Anderson H.R., Sunyer J., Ayres J., Baccini M., Vonk J.M., Boumghar A., Forastiere F., Forsberg B., Touloumi G., Schwartz J., Katsouyanni K. (2001). Acute effects of particulate air pollution on respiratory admissions: results from APHEA 2 project. Air Pollution and Health: a European Approach. *Am J Respir Crit Care Med,* **15**, 164(10 Pt 1), 1860–1866.

Biggeri, A., Bellini, P., Terracini, B. (2001). Meta-analysis of the Italian studies on short-term effects of air pollution, *Epidemiol. Prev.*, **25**(2 Suppl):1-71. In Italian.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., (1994). *Time Series Analysis: Forecasting and Control*. Prentice-Hall Inc.

Clyde, M. (2000). Model uncertainty and health effect studies for particulate matter, *Environmetrics*, **11**, 745–763.

Chiogna, M., Gaetan, C.,Masarotto, G. (2003). Automatic identification of Transfer Function models by means of iterative stepwise and genetic algorithms. (submitted)

Dominici F., Zeger S.L., and Samet J.M. (2000). Combining Evidence on Air Pollution and Daily Mortality from the Largest 20 US cities: A Hierarchical Modeling Strategy (with discussion). *Journal of the Royal Statistical Society, Series A*, **163**, 263-302.

Draper, N., Smith, H., (1998). *Applied Regression Analysis*, 3rd Edition. Wiley, New York.

Gaetan, C., 2000. Subset ARMA model identification using genetic algorithms. *Journal of Time Series Analysis*, **21**, 559–570.

Hannan, E. J., Quinn, B. G., (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, **41**, 190–195.

Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T. (1999). Reply to comments on "Bayesian Model Averaging: A Tutorial ", *Statistical Science*, **14**, 412–417.

Holland, J., (1975). *Adaption in Natural and Artificial systems: an Introductory Analysis with Applications to Biology, Control and AI*. The University Michingan Press, Ann Arbor, MI.

Miller, A. (1990). *Subset Selection in Regression*, Chapman and Hall, London.

Samet, J.M., Dominici, F., Zeger, S.L., Schwartz, J., Dockery, D.W. (2000). The National Morbidity, Mortality, and Air Pollution Study Part I: Methods and Methodologic Issues. Research Report Number 94, The Health Effects Institute, Cambridge, MA.

Samet, J.M., Zeger, S.L., Dominici, F., Curriero, F., Coursac, I., Dockery, D.W. Schwartz, J., Zanobetti, A. (2000). The National Morbidity, Mortality, and Air Pollution Study Part II: Morbidity and Mortality from Air Pollution in the United States, Research Report Number 94 Part II, The Health Effects Institute, Cambridge, MA.

Normand, S.L. (1999). Meta analysis: Formulating, evaluating, synthesizing, and reporting. *Statistics in Medicine*, **18**, 321–359.

Schwartz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, 461–464.

Smith, R., Davis, J., Sacks, J., Speckman, P. Styer, P. (2000). Regression models for air pollution and daily mortality: analysis of data from Birmingham, Alabama, *Environmetrics*, **11**, 719–743.

Whitley, D. (1989). The GENITOR algorithm and selective pressure, *Proc 3rd International Conference on Genetic Algorithms*, Morgan Kaufmann, 116–121.