

The estimation of standardized trends of pollutants using hierarchical models ⁽¹⁾

*Modelli gerarchici per la stima del trend standardizzato di inquinanti
atmosferici*

Enrico Fabrizi, Carlo Trivisano

Dipartimento di Scienze Statistiche “Paolo Fortunati”

Università di Bologna

fabrizi@stat.unibo.it, trivi@stat.unibo.it

Riassunto: In questo lavoro si propongono, comparandole, tre diverse classi di modelli per la stima del trend dei massimi giornalieri delle concentrazioni di ozono rilevati nell’area urbana di Bologna nel periodo 1994-98. Al fine di eliminare l’effetto confondente della meteorologia sul trend è stato seguito un approccio di stratificazione che utilizza gli alberi di regressione. Si sono confrontati tra loro modelli a effetti casuali basati sulle ipotesi di distribuzione Weibull e Normale. I risultati mostrano che è preferibile assumere l’eteroschedasticità delle distribuzioni specifiche di ciascun stato e che vi è una insufficiente evidenza empirica a sostegno dell’ipotesi di trend diversi per differenti livelli del processo.

Keywords: Ozone trends, Hierarchical models, Random effects, MCMC methods

1. Introduction

This paper analyzes the trend of maximum daily levels of surface ozone concentrations from a monitoring site located in the urban area of Bologna. A major issue in the analysis of ozone trends is to assess whether the observed trend in the data is due to variations in the emissions or to changing meteorological conditions.

To take the ozone dependence on meteorology into account, we consider a stratified approach (Thompson *et al.*, 2000), according to the proposal of Huang and Smith (1999). In the models introduced by these two authors, homogeneous meteorological regimes are identified by means of regression tree models. Tree based stratified models enjoy two main advantages: they can manage complex relationships between ozone concentrations and meteorological variables and permit the verification of the existence of different trends under different meteorological regimes. For trend determination Huang and Smith (1999) introduce a set of alternative linear homoschedastic normal random effects models, some of which allows for specific trends for each group. Cocchi *et al.* (2001) propose a different stratified model in which maxima are assumed to be Weibull distributed within each group identified by the tree, with a varying scale but a common shape parameter. The Weibull distribution assumption arises naturally in the context of maxima; moreover it represents a flexible choice since, for different values of its parameters, it ranges from approximately normal to highly skewed forms. On the other hand, using a common shape parameter implies constant coefficient of variation

⁽¹⁾ The research leading to this paper has been partially funded by 2000 grant (project No 26, sector: Economics and Statistics) for research projects of national interest by MIUR

across groups and thus a heteroschedastic model. Cocchi *et al.* (2001) introduce a random effects model for the logarithm of the quasi-scale parameter of the Weibull distribution, where the considered effects are represented by the homogeneous ozone regimes resulting from tree partitioning and the year. The trend is then defined as the sequence of the parameters associated to year effects. The models based on the Weibull distribution are shown to perform well for the same Bologna data set we consider here. Nonetheless, the form of the group-specific Weibull distributions turn out to be generally close to that of the normal.

In this paper, we apply the assumption of constant coefficient of variation to a normal stratified model and compare it to the analogous model based on the Weibull distribution. Our aim is to assess whether the good performances achieved by the model proposed by Cocchi *et al.* (2001) are due to the relaxing of the homoschedasticity assumption adopted by Huang and Smith or to their different distributional hypotheses.

As regards inference, we adopt a full Bayesian approach. Models are solved by means of Gibbs sampling routines, as they are implemented in the software BUGS (Spiegelhalter *et al.*, 1996).

2. The data set

We analyzed the series of daily maxima of ozone concentrations over the metropolitan area of Bologna in the period 1994-1998. Maxima are calculated on the basis of hourly measurements from a single monitoring site situated in Giardini Margherita, a park close to the center of the city. We consider exclusively data gathered during the ozone seasons (from April 1st to October 31st) yielding a gross total of 1,067 observations. For 66 observations the maximum is not calculated because of measurement failures or for other causes. The meteorological variables we make use of are collected at the surface weather station of Borgo Panigale, a few kilometers away from the monitoring site of Giardini Margherita. Among the available measurements, we consider the noon values of temperature (°C), dew point temperature (°C), pressure (*mb*), relative humidity (%), visibility (*km*), wind speed (*m/s*) and wind direction (° from North).

3. The models

We built a regression tree for ozone daily maxima using the set of meteorological covariates described in section 2. We adopted the CART approach (Breiman *et al.*, 1984), as it is free of distributional assumptions. The split criterion is given by a square loss function and the usual cost-complexity criterion described in Breiman *et al.* (1984) is considered for pruning the tree; cross-validation is employed to select the degree of pruning. On the basis of the regression tree we identify eight different homogeneous ozone regimes. As regards the models for trend determination, in both the Normal and Weibull distributional context, we propose to model the median of observations as a function of random effects, represented by tree-based groups and year intercepts. For the Normal case we specify an additive model, since the median is a location parameter, while in the Weibull case we suggest a multiplicative model since the median is proportional to the scale parameter (we parameterize the Weibull as in Evans *et al.*,

1993). Let's now depict the two competing models for the cases characterized by a common trend. In the Weibull model we have

$$y_i | \lambda_i, \zeta \stackrel{ind}{\sim} Weib(\lambda_i, \zeta) \quad i = 1, \dots, 1067 \quad (1)$$

We may then write that

$$\log(\mathbf{M}) = \mathbf{1} + \mathbf{A} + \mathbf{B} \quad (2)$$

where the vector $\mathbf{1}$ indicates the general intercept, while $\mathbf{A}$ and $\mathbf{B}$ are the design matrixes for the group and year effects respectively. The vector $\mathbf{M}$ contains individual medians, each defined as

$$M_i = Med(y_i) = (\log 2)^{1/2} \lambda_i \quad (3)$$

The normal model can be described by

$$y_i | \mu_i, \theta \stackrel{ind}{\sim} Norm(\mu_i, \mu_i^2 / \theta) \quad (4)$$

$$\boldsymbol{\mu} = \mathbf{1} + \mathbf{A} + \mathbf{B} \quad (5)$$

As regards the specification of prior distributions, we assume, in the Weibull model, a $Gamma(3,1)$ for the shape parameter ζ implying $E(\zeta) = 3$ and $V(\zeta) = 3$. This distribution is centered on the value of ζ for which the Weibull has a shape similar to that of the Normal, but is sufficiently diffuse to give support also to the hypotheses of more skewed distributions. On the contrary, for the parameter θ in the Normal model we select a diffuse $Gamma(0.01, 0.01)$. For all the parameters related to the distributions of random effects we specify similar hierarchical priors in both classes of models, as described in Cocchi *et al.* (2001). Models with group-specific trend can be written simply replacing $\mathbf{1} + \mathbf{A} + \mathbf{B}$ with the matrix $\mathbf{C}$ for the interactions between year and group effects, while the models without trends are individuated by the only general intercept $\mathbf{1}$ in (2) and (5).

4. Model comparisons and discussion of empirical results

Models are compared in terms of ΔBIC (Schwartz, 1978) that represents an asymptotic, prior assumption free, approximation of the Bayes Factor. In table 1 we compare models assuming respectively common, group-specific and no trends, specified under the three different distributional assumptions we considered.

There is evidence that models with common trends for all the ozone regimes are best regardless of the modeling assumptions considered.

Table 1: ΔBIC for the comparison of models with different distributional assumptions

	Weibull	Eterosch. Normal	Omosch. Normal
Specific vs. Common trend	-151.3	-156.0	-144.4
Specific vs. No trend	-1.3	-5.0	9.8
Common vs. No trend	150.0	151.0	154.2

In table 2 we compare the homoschedastic Huang and Smith model with the heteroschedastic Weibull and normal models described in section 3.

Table 2: ΔBIC for the comparison of models with different trends assumptions

	No trend	Common trend	Specific trend
Eterosch. Normal vs. Weibull	18.9	19.9	15.2
Eterosch. vs. Omosch. Normal	42.1	38.9	27.3
Weibull vs. Omosch. Normal	23.2	19.0	12.1

The model comparisons highlight that the heteroschedastic normal models perform better than the homoschedastic ones and than those based on the Weibull distribution. At least for the data set we consider, there is then evidence that heteroschedasticity needs to be taken into account. Moreover, even though the posterior distribution of μ_i in (4) cannot be calculated in closed form, we found, from MCMC simulations that they can have positive asymmetry, thus making them able to fit the moderately asymmetric distribution of maxima within groups very well. We have also checked the robustness of these results with respect to the choice of the splitting criterion in the regression tree, replacing the quadratic loss described in section 3 with the absolute loss. The tree identifies, in this case, only seven groups, but comparisons among models for trend determination remain largely unchanged.

References

- Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. (1984) *Classification and regression trees*, Wadsworth, California
- Cocchi D., Fabrizi E., Trivisano C. (2001) *A stratified model for the analysis of ozone trends in urban area*, Working paper, Dipartimento di Scienze Statistiche “Paolo Fortunati”, Università di Bologna
- Evans M., Hastings N., Peacock (1993) *Statistical distributions*, Wiley, New York
- Huang, L.-S. and Smith, R.L. (1999). Meteorologically-dependent trends in urban ozone, *Environmetrics*, 10, 103-118.
- Schwartz G. (1978) Estimating the dimension of a model, *Annals of Statistics*, 6 461-64
- Spiegelhalter D.J., Thomas A., Best N. AND Gilks W.R. (1996) BUGS: *Bayesian Inference Using Gibbs Sampling, version 0.50*, Technical Report, Medical Research Council Biostatistic Unit, Institute of Public Health, Cambridge University
- Thompson, M.L., Reynolds, J., Cox, L.H., Guttorp, P. and Sampson, P.D. (2000). *A review of statistical methods for the meteorological adjustment of tropospheric ozone (revised)*, Technical Report Series #26, NRCSE-TRS.