ISSN 1032-3813
ISBN 0 7326 1039 7

MONASH UNIVERSITY

Estimating long-term trends in tropospheric ozone levels

Michael Smith, Paul Yau, Thomas Shively and Robert Kohn

Working Paper 2/98
April 1998

DEPARTMENT OF ECONOMETRICS
AND BUSINESS STATISTICS

Estimating long-term trends in tropospheric ozone levels.

Michael Smith[a], Paul Yau[b], Thomas Shively[c] and Robert Kohn[b]

[a] Department of Econometrics and Business Statistics, Monash University

[b] Australian Graduate School of Management, University of New South Wales

[c] Department of Management Science and Information Systems, University of Texas at Austin

April 1, 1998

Abstract

This paper estimates the long-term trends in the daily maxima of tropospheric ozone at six sites around the state of Texas. The statistical methodology we use controls for the effects of meteorological variables because it is known that variables such as temperature, wind speed and humidity substantially affect the formation of tropospheric ozone. A nonparametric regression model is estimated in which a general trivariate surface is used to model the relationship between ozone and these meteorological variables because there is little, or no, theory to specify the functional dependence of ozone on these variables. The model also allows for the effects of wind direction and seasonality. Each function in the model is represented as a linear combination of basis functions located at all of the design points. A trivariate basis is used for the function representing the combined effect of temperature, wind speed and humidity, while univariate bases are used to represent the other functions in the model. To estimate the functions nonparametrically we use a Bayesian hierarchical framework with a fractional prior. Due to the high dimensional representation of the signal, a Markov chain Monte Carlo sampling scheme employing Gibbs sub-chains that 'focus' on the basis terms that are most likely to contribute to the signal is used to carry out the computations. We

also estimate an appropriate data transformation simultaneously with the function estimates. The empirical results indicate that key meteorological variables explain most of the variation in daily ozone maxima through a nonlinear interaction and that their effects are consistent across the six sites. However, the estimated trends vary considerably from site to site, even within the same city. A simulation based on the design of the data indicates that the Bayesian approach is substantially more efficient than MARS (Friedman, 1991).

Key Words: Data transformation; Focused sampling; Nonparametric regression; Reproducing kernel; Trivariate Radial Basis

[1]Both Michael Smith and Robert Kohn's work was partially supported by funds from the Australian Research Council. The authors would like to thank the staff at the Texas Natural Resource Conservation Commission for providing the data and many useful discussions with regard to the empirical results. In this regard, we would especially like to thank M.W. Hemphill, Bryan Lambeth and Larry Butts. We would also like to thank Professor J. Friedman for providing us with his Fortran 77 code implementing MARS.

1 Introduction

A major issue with the analysis of data on tropospheric ozone is to establish whether observed trends can be attributed to the effects of pollution control programs implemented over the past two decades, or whether they are the result of meteorological changes affecting the conditions under which ozone is generated. Tropospheric ozone refers to ozone in the ambient air, not ozone in the upper atmosphere. Ozone in the ambient air is an air pollutant and can have a significant impact on people's health, particularly in children, the elderly and those with lung disease. Therefore, one would like to see a downward trend through time in tropospheric ozone levels.

The formation of ozone results from a chemical reaction in the ambient air involving nitrogen oxides and volatile organic compounds. The chemical reaction that produces ozone is complex and not completely understood, even in the laboratory. However, it is known that the reaction is largely driven by a combination of key meteorological conditions in what is likely to be a nonlinear manner. Therefore, even if pollution control programs are successful in reducing the emissions of toxic gases into the atmosphere, a downward trend may not be observed in the raw ozone data due to the effects of changing meteorological conditions. Such conditions should be taken into account to obtain a reliable estimate of the long-term trend in daily ozone levels.

This paper uses a Bayesian approach to estimate a nonparametric regression model for observations of daily tropospheric ozone maxima at six monitoring stations in Texas during the period 1980-1997. The model incorporates the combined effect of the key variables of wind speed, temperature range (which acts as a proxy for sunlight) and humidity as a nonparametric trivariate interaction surface. The effects of wind direction, seasonal and trend variables are accounted for as additive univariate nonparametric functions. Each of the functions are modeled as linear combinations of basis terms, with locations at all the unique design points. A wide variety of basis expansions can be employed. We use a trivariate radial basis to represent the function relating ozone to wind speed, temperature range, and humidity; univariate reproducing kernels as the basis functions for the univariate functions relating ozone

to wind direction; a dummy variable basis to represent the function modeling seasonality; and a linear regression spline for the trend function. To estimate the regression coefficients we use an adaptation of the hierarchical Bayesian model initially discussed in Smith and Kohn (1996), coupled with a fractional prior of the type discussed by O'Hagan (1995). To deal with the high dimensional basis representation of the regression functions an adaption of the focused sampling scheme introduced in Wong, Hansen, Kohn and Smith (1997) is used for the computations. As the empirical work here demonstrates, the resulting estimator is both automatic and applicable to complex multiple nonparametric regressions with large sample sizes.

There have been several recent studies of tropospheric ozone. For example, Nychka, Yang and Royle (1998) discuss optimal location of monitoring sites in the Chicago urban area for spatial models of ambient air ozone, but are not concerned with identifying long-term trends or the role of meteorological variation. Carroll, Chen, George, Newton, Schmiediche and Wang (1997) also develop a spatial model for twelve monitoring sites in Harris County, Texas. Their analysis examines a global trend for the county, but does not consider local site-based trends, nor take account of the complex nonlinear relationship between key meteorological variables and ozone levels. Smith and Huang (1993) and Shively (1990) analyzed exceedences of legislative thresholds for tropospheric ozone using extreme value theory. However, following Cox and Chu (1992), Bloomfield, Royle and Yang (1993) and Niu (1996) we examine daily ozone maxima. This provides a better understanding of the trends in long-term (chronic) exposure to relatively low levels of ozone than threshold exceedences; an issue that is of keen interest to the Texas Natural Resource Conservation Commission, who collected the data in this study. Figure 1 provides boxplots of the daily maxima for the Aldine monitoring site, indicating that a data transformation may be required to ensure that a Gaussian model for the errors is appropriate. Therefore, we estimate a data transformation from a discrete set of potential Box-Cox style power transformations simultaneously with the unknown functions. These transformations are normalized to be location and scale invariant to make it easier to interpret the empirical results.

—Figure 1 About Here.—

Other authors have also accounted for meteorological variation in tropospheric ozone. Bloomfield et al. (1993) control for a large number of meteorological variables using a two stage procedure. First, they use 'loess' (Cleveland, Grosse and Shyu, 1992) to suggest appropriate parametric functional forms for the bivariate relationships between (i) ozone, temperature and wind speed; and (ii) ozone, temperature and humidity. These are then included in a nonlinear parametric regression. It is difficult to obtain reliable function estimates using this approach because loess relies on a subjective exploratory approach to determine an appropriate smoothing parameter, while the two stage procedure can induce a mis-specification problem because each of the functional forms determined in the first stage are obtained without controlling for the other independent variables. Alternatively, Smith and Huang (1993) account for an interaction effect between temperature and wind speed by using a parametric model with the multiplication of temperature and wind speed as an independent variable.

Rather than pre-determine a parametric regression model, Niu (1996) develops an additive nonparametric model in the meteorological variables, where the functional relationships are estimated from the data. He adapts a back-fitting algorithm to estimate all the functions, while also estimating a parametric time series model for the error terms. Smoothing splines are used as the univariate smoothers, with smoothing parameters estimated using generalized cross-validation. However, efficient determination of the smoothing parameters that drive each of the underlying smoothers is often difficult with the mis-specification of any single parameter possibly resulting in poor estimates for all component functions. Importantly, the model is an additive model and no interaction effects between key variables are considered. Similarly, Shively and Sager (1997) use an additive model of univariate smoothing splines (Wahba, 1990). To attempt to account for interactions some pairwise multiplications of the meteorological variables, as well as the meteorological variables themselves, are included as regressors. However, it is not clear that such an additive structure is appropriate and secondly, no attempt to account for three way interactions is undertaken.

In comparison to previous work, our procedure does not require the explicit estimation

of smoothing parameters and can easily incorporate full nonparametric interaction surfaces through the use of an appropriate basis, such as the trivariate radial basis in wind speed, temperature range and humidity introduced in section 3. Our empirical work suggests that daily ozone maxima are greatly affected by such interactions. Few alternative data-driven methodologies exist that can estimate high dimensional nonparametric regression models with interaction surfaces and higher sample sizes. For example, tensor product multivariate smoothing splines (Gu, Bates, Chen and Wahba, 1989) are $O(n^3)$ and are computationally infeasible for the large sample sizes used here. While local regression based techniques theoretically also extend to such multivariate models, estimation of the bandwidth parameter(s) is also computationally infeasible. One viable alternative is MARS (Friedman, 1991) which uses a search algorithm on tensor product regression splines. To assess our empirical results, a simulation is performed that generate data from both our fitted model and that resulting from a MARS fit to the same regression model. We show that in both cases the Bayesian approach is better than MARS at reproducing the true models.

The paper is organized as follows. Section 2 contains a description of the data analyzed in the paper. Section 3 describes the nonparametric regression used to model the ozone data, including the bases used to model each of the functions. Section 4 discusses how such a model can be interpreted in a Bayesian hierarchical framework and develops the 'focused' Markov chain Monte Carlo sampling scheme used to undertake the computation. The empirical results are presented and discussed in section 5. The simulation comparison with MARS is undertaken in section 6, while section 7 contains some conclusions.

2 The Data

The data used in this paper were collected at six Texas monitoring sites and provided to us by the Texas Natural Resource Conservation Commission (TNRCC). Figure 2 provides a map showing the location of the sites. The Aldine, Clinton and Northwest Houston sites are located in Houston, the Fort Worth Keller and Dallas North sites are located in the Dallas-Fort Worth Metroplex area, while the final site is located at Beaumont. These sites

are of particular interest to the TRNCC as they represent the two major metropolitan areas of Texas and a major industrial area (Beaumont).

——Figure 2 About Here.——

The data consists of daily maximum ozone values observed at these sites during the months May-October over the eighteen year period 1980-1997. The months May-October are considered the "high ozone" season and is the time of the year when ozone in the ambient air typically creates a problem. Also collected at each site were daily values of important meteorological variables. The variables we use in our analysis are given below.

- Ozone (OZ): The daily ozone value used in this study is the maximum of the 13 hourly ozone readings (in parts per hundred million) taken each hour from 6am to 6pm.
- Temperature range (TR): Difference between the minimum and maximum hourly temperature readings for the period 6am to 6pm. The temperature range is a well-accepted proxy for the amount of sunlight occurring during the day because the temperature range increases as the amount of sunlight increases. (A direct measure of sunlight is not available at the monitoring sites). The expected relationship between temperature range and ozone levels is positive.
- Wind speed (WS): Average of the hourly wind speed readings for the period 6am to 6pm. The expected effect of increased wind speed is to reduce ozone levels because higher wind speed tends to disperse pollutants present in the ambient air.

The datasets also include four wind direction variables measuring the proportion of time between 6am and 6pm when the hourly wind direction fell into one of four 90 degree quadrants. These quadrants differ from site to site and they are defined in table 1. We define WD_1, WD_2, WD_3 and WD_4 to be the percentage of time from 6am and 6pm that the wind direction fell into each of these four quadrants. Because these variables sum to one, we only include WD_2, WD_3 and WD_4 into our analysis.

—Table 1 About Here.—

Two other variables are also used in our regression model and are:

- Monthly variable (MN): Here, $MN = 5, 6, 7, 8, 9$ or 10 if the observation occurs in May, June, July, August, September or October, respectively. This variable is used to model seasonality in the ozone data during the high ozone season, over and above that captured by the meteorological variables above.
- Annual trend term (YR): $YR_t = 1, 2, \ldots, 18$ if day t is in $1980, 1984, \ldots, 1997$, respectively. This variable is used to model the long-term trend in ozone values.

The following missing data convention is used for the ozone and meteorological data. If more than 7 hourly readings in the period 6am to 6pm are missing on a given day for the ozone or for any meteorological variable, then the data for that day are considered to be missing. Table 2 outlines the years during which data from each station were collected, along with the number of observations and percentage of missing data.

—Table 2 About Here.—

3 The Nonparametric Regression Model

We model daily ozone maxima at each of the six sites with the nonparametric regression model

$$\begin{aligned} T_\lambda(OZ_i) &= \alpha + f_1(TR_i, WS_i, HMD_i) + f_2(WD_{1,i}) + f_3(WD_{2,i}) \\ &+ f_4(WD_{3,i}) + f_5(MN_i) + f_6(YR_i) + e_i\,. \end{aligned} \tag{3.1}$$

Here, f_1 is a smooth, but unknown, trivariate function that models the interaction effect of temperature range, wind speed and humidity. The wind direction effects enter the model

additively as nonparametric univariate functions f_2, f_3 and f_4. Any seasonal effect over and above that pertaining to the meteorological variables, is captured by f_5. The function f_6 measures the long-term trend in ozone, controlling for the effect of meteorological conditions and seasonality.

Figure 1 highlights the highly skewed distribution of daily maxima of hourly tropospheric ozone values. Previous authors consider various Box-Cox style data transformations, but do not attempt to estimate such transformations in combination with the signal. Therefore, we estimate the most appropriate transformation simultaneously with the unknown functions in the regression model at (3.1). We consider a location and scale invariant transformation $T_\lambda(OZ)$, indexed by λ, of the form

$$T_\lambda(OZ) = a_\lambda + b_\lambda t_\lambda(OZ)$$

where

$$t_\lambda(OZ) = \begin{cases} (OZ+1)^\lambda & \text{if } \lambda > 0 \\ \log(OZ+1) & \text{if } \lambda = 0 \\ -(OZ+1)^\lambda & \text{if } \lambda < 0 \end{cases}$$

for the discrete set of values of $\lambda \in \Lambda = \{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1.0\}$. The 'base' transformation t_λ is a monotonic Box-Cox style power transformation where we add one to OZ because $\min_i(OZ_i) = 0$. For the data collected from each monitoring site, this transformation is then normalized by constants a_λ and b_λ to produce the data transformation T_λ. These constants are calculated as in Smith and Kohn (1996) so that the data have approximately the same median and inter-quartile range before and after transformation. This normalized transformation is used because it does not alter the scale or location of the data and therefore eases the qualitative interpretation of the regression results.

Each of the unknown functions in the regression at (3.1) is modeled as a linear combination of basis functions, so that for any point z in the domain of the independent variable,

$$f_j(z) = \sum_i \beta_i^j b_i^j(z) \quad \text{for } j = 1, 2, \ldots, 6\,.$$

The β_i^j are coefficients requiring estimation and the $b_i^j \in \mathcal{B}^j$ are basis functions located at