MONASH

WP 2/98

ISSN 1032-3813
ISBN 0 7326 1039 7

MONASH UNIVERSITY

ANCORA · IMPARO

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 every unique design point. The type of each of the bases $\mathcal{B}^1, \mathcal{B}^2, \ldots, \mathcal{B}^6$ are chosen according to the nature of the effect and are listed below.

- For the trivariate function we use a radial basis (Powell, 1987; Holmes and Mallick, 1997), with $\mathcal{B}^1 = \{z_1, z_2, z_3, ||\boldsymbol{z} - \boldsymbol{x}_1||^2 \log(||\boldsymbol{z} - \boldsymbol{x}_1||), \ldots, ||\boldsymbol{z} - \boldsymbol{x}_n||^2 \log(||\boldsymbol{z} - \boldsymbol{x}_n||)\}$. Here, $\boldsymbol{x}_i = (TR_i, WS_i, HMD_i)'$ and $\boldsymbol{z} = (z_1, z_2, z_3)'$.

- For the univariate wind direction functions f_2, f_3 and f_4 we used a reproducing kernel basis (Luo and Wahha, 1997) which Wong et al. (1997) demonstrate is a good basis for the estimation of smooth univariate functions. It is defined for each wind direction as $\mathcal{B}^j = \{z, R(z, x_1), \ldots, R(z, x_n)\}$, where $x_i = WD_{j,i}$ and
$$R(z, x_i) = \left(-(|z - x_i| - \frac{1}{2})^4 + \frac{1}{2}(|z - x_i| - \frac{1}{2})^2 - \frac{7}{240} \right) /24$$
The reproducing kernel basis is defined over the unit interval, so the wind direction independent variables are scaled to $[0, 1]$ upon calculation of the terms, though the results are interpreted on the original scale.

- The seasonal component f_5 is modeled using a dummy variable basis which is intended to capture significant monthly deviations in ozone maxima from the mean α. The basis is defined as $\mathcal{B}^5 = \{I(z, x_1), \ldots, I(z, x_n)\}$ where $x_i = MN_i$ and
$$I(z, x_i) = \begin{cases} 1 & \text{if } z = x_i \\ 0 & \text{otherwise} \end{cases} .$$

- The trend component f_6 uses a linear regression spline because the basis terms are ramp functions that capture any significant alterations in the trend. The basis is $\mathcal{B}^6 = \{z, (z - x_1)_+, \ldots, (z - x_n)_+\}$ where $x_i = YR_i$ and $(x)_+ = \max(0, x)$.

If there are no replicated values, each of these bases would contain approximately n terms. However, in our datasets there are large number of replicated values for the independent variables, especially for the wind directions, month and year. Table 3 details the number of basis terms for each function and the resulting dimension of the basis representation of the signal for the data arising from each monitoring site.

—Table 3 About Here.—

4 Methodology

4.1 Hierarchical Bayesian Model

Given these basis terms and the index of the data transformation, λ, the regression for a particular site is simply a parametric linear model of the form

$$y_\lambda = X\beta + e \tag{4.1}$$

Here, the $(n \times p)$ design matrix X is made up of all the basis terms introduced above, along with a column of 1's for the global intercept α. The p-vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ contains the regression coefficients, $e = (e_1, \ldots, e_n)' \sim N(0, \sigma^2 I_n)$ are errors and $y_\lambda = (T_\lambda(OZ_1), \ldots, T_\lambda(OZ_n))'$ is the vector of dependent variable values.

One way to render this regression nonparametric is to estimate the regression parameters β (and hence the unknown functions in equation (3.1)) using a Bayesian hierarchical model. This model was discussed in Smith and Kohn (1996) and explicitly accounts for the uncertainty that each term will enter the regression; a brief exposition is given below.

Let γ be a p-vector of indicator variables with the ith element γ_i, such that $\gamma_i = 0$ if $\beta_i = 0$ and $\gamma_i = 1$ if $\beta_i \neq 0$. Given γ, let β_γ consist of all the nonzero elements of β and let X_γ be the columns of X corresponding to those elements of γ that are equal to one. Therefore, the linear model can be rewritten conditional on γ as

$$y_\lambda = X_\gamma \beta_\gamma + e$$

To form the hierarchical model, the following prior assumptions are made on the model parameters:

A1: Following O'Hagan (1995) we take a fractional conditional prior

$$p(\beta_\gamma | \gamma, \sigma^2, \lambda) \propto p(y_\lambda | \beta_\gamma, \gamma, \sigma^2, \lambda)^{1/n},$$

so that

$$\beta_\gamma|\gamma, \sigma^2, \lambda \sim N\left(\mu_\gamma, n\sigma^2(X'_\gamma X_\gamma)^{-1}\right),$$

where $\mu_\gamma = (X'_\gamma X_\gamma)^{-1}X'_\gamma y_\lambda$. This provides little information on the location of β_γ compared to the likelihood because the conditional prior variance is scaled up by a factor of n.

A2: The prior of σ^2 is taken *a priori* independent of γ and λ, so that $p(\sigma^2|\gamma, \lambda) \propto 1/\sigma^2$. This is a commonly used prior for σ^2 because it makes $\log(\sigma^2)$ uniform.

A3: The prior for γ is taken *a priori* independent of λ and the elements are independent and identically distributed $p(\gamma_i|\gamma_{j\neq i}) = 1/2$. This ensures that the prior $p(\gamma) = 2^{-p}$.

A4: All nine potential transformations are assumed equally likely a priori, with $p(\lambda = i) = 1/9$ for $i = 1, 2, \ldots, 9$.

Using these priors, the posterior probability of any particular subset of variables γ can be calculated as

$$p(\gamma|y) \propto (n+1)^{-q_\gamma/2} \sum_{\lambda\in\Lambda} S(\gamma, \lambda)^{-n/2} J_\lambda.$$

Here, $S(\gamma, \lambda) = y'_\lambda y_\lambda - y'_\lambda X_\gamma(X'_\gamma X_\gamma)^{-1}X'_\gamma y_\lambda$, J_λ is the Jacobian of the data transformation (see Smith and Kohn (1996) for details on its calculation) and $q_\gamma = \sum_{i=1}^p \gamma_i$ is the number of non-zero regression coefficients in model γ. This is almost the Schwarz (1978) information criteria for a particular subset of regression terms in the linear model at (4.1). The problem here is that γ has support on 2^p possible subsets and due to the basis representation of the signal is a formidable nuisance parameter in estimating both the regression coefficients β and transformation $\lambda \in \Lambda$ in the hierarchical linear model (4.1).

4.2 Focused Sampling Scheme

Because of the high dimension of the basis representation of the signal, estimating the regression and transformation parameters in the hierarchical model using the 'one at a time' Gibbs sampling scheme discussed in Smith and Kohn (1996) would be computationally burdensome. Alternative sampling schemes include the reversible jump sampler (Green, 1995) which has been applied to univariate nonparametric regression models by Denison, Mallick and Smith (1998) and radial basis functions by Holmes and Mallick (1997). However, to solve this problem we use the following generalization of the Gibbs sampling scheme discussed in Smith and Kohn (1996) and focused sampling steps discussed in Wong, et al., (1997).

Focused Sampling Scheme.

```
Step (0): Select initial state $\gamma = \gamma^{[0]}$
Step (1): Choose $\mathcal{S} \subset \{1, 2, \dots, p\}$ in a probabilistic manner
Step (2): Repeat the following $M$ times:
          Sequentially generate $\gamma_i | \gamma_{j \neq i}, \boldsymbol{y}$ for $i \in \mathcal{S}$
```

The selection of $\mathcal{S}$ in Step (1) is performed so that $Pr(i \in \mathcal{S}) > 0$ for $i = 1, 2, \dots, p$. In this case the resulting Markov chain is irreducible and aperiodic and therefore converges to its invariant distribution, which is $\gamma | \boldsymbol{y}$ (Tierney, 1994).

The scheme is run in three stages. The first is a warmup period of length 2000 iterations, after which the sampler is assumed to have converged. The second is a sampling period of length $K_1 = 1000$ iterations in which the distribution of $\lambda | \boldsymbol{y}$ is estimated and from which the mode estimate $\hat{\lambda}$ is obtained. The third sampling period is of length $K_2 = 4000$ and conditions on this mode transformation, so that the distribution from which generation is undertaken in Step (2) is now $\gamma_i | \gamma_{j \neq i}, \lambda = \hat{\lambda}, \boldsymbol{y}$. We condition on such a single best transformation, rather than smooth over its distribution, to make analysis of the results simpler as advocated in Box and Cox (1982).

Step (2) is a Gibbs sub-chain of length M that converges to the conditional posterior distribution of the subvector $\gamma_{i \in \mathcal{S}} | \gamma_{i \notin \mathcal{S}}, \boldsymbol{y}$. Note that if $\mathcal{S} = \{1, 2, \dots, p\}$ in Step (1) for every iteration of the sampler, then this scheme simply reduces to Gibbs sampling. However, in our problem p is large, so to both reduce the number of generations required and reduce the dependence among Markov chain iterates we use the following at Step (1), for any iteration k, to adaptively focus on basis terms that are more likely to be important.

```
Step (1) at iteration $k$
(1a) $\mathcal{S}_1 = \{i|\gamma_i^{[k-1]} = 1\}$; $\mathcal{S}_2 = \emptyset$
(1b) do $i = 1,..,6$
    (1b-i) $P = \max(\frac{q_i}{p_i - q_i}, 0.05)$
    (1b-ii) $P = \min(1, P)$
    (1b-iii) Add the indices of the each of the basis terms associated
    with function $f_i$ to $\mathcal{S}_2$ with probability $P$.
(1c) $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$
```

Here, p_i is the number of basis terms arising from function f_i (see table 3), while q_i is the number of basis terms actually selected as non-zero for f_i in the previous $(k-1)$th iteration. These steps ensure that the binary variables for those terms that were significant last iteration are always generated again, labeling them with the index set $\mathcal{S}_1$. We choose a subset of the remaining binary variables, labeled with the indexing set $\mathcal{S}_2$, using the following probabilistic rule. For each of the functions, we select binary variables to generate with a probability that, on average, ensures at least 5% of the binary variables for each function are generated. However, if a function is highly oscillatory and requires more basis terms in the previous iteration, then a large number of additional indices are selected for generation. This ensures that (i) generation of terms is dynamically allocated to those binary variables whose corresponding basis terms are likely to be significant. (ii) Despite the wide discrepancy in the number of basis terms for each component signal in the regression model at all six sites (see table 3) equal focus can be maintained on each component function.

This scheme is computationally efficient, compared to the equivalent full Gibbs sampler, in that time is not allocated to repeatedly generate the majority of the many binary variables that are unlikely to have a significant impact on the function estimates. We use the iterates $\{\gamma^{[1]}, \ldots, \gamma^{[K_1]}\}$ to calculate mixture estimates for the posterior distribution $p(\lambda|\boldsymbol{y})$ with

$$\hat{p}(\lambda|\boldsymbol{y}) = \frac{1}{K_1} \sum_{k=1}^{K_1} p(\lambda|\gamma^{[k]}, y) \quad \text{and } \hat{\lambda} = \text{argmax}_{\lambda \in \Lambda} \hat{p}(\lambda|\boldsymbol{y})$$

The iterates $\{\gamma^{[K_1+1]},\ldots,\gamma^{[K_2]}\}$ are used to estimate the posterior function means $E[f_i(z)|y,\lambda=\hat{\lambda}]$ with the mixture estimate

$$\hat{f}_i(z) = x^{i\prime}\hat{\beta}^i \quad \text{where } \hat{\beta} = \frac{1}{K_2}\sum_{k=K_1+1}^{K_2} E[\beta|\gamma^{[k]},\lambda=\hat{\lambda},y] \tag{4.2}$$

Here, $x^{i\prime}$ is the p_i-vector of terms arising from calculating the basis terms for f_i at point z. The vector $\hat{\beta}^i$ is a subset of the elements of $\hat{\beta}$ that correspond to basis terms for f_i. The conditional expectation at (4.2) can be calculated exactly as $\beta_j = 0|\gamma_j = 0$ and $E[\beta_\gamma|\gamma,\lambda,y] = (X_\gamma' X_\gamma)^{-1}X_\gamma' y_\lambda$.

The estimate of the posterior standard deviation of f_j evaluated at a point z is calculated as

$$\hat{s}_j(z) = \left(\frac{1}{K_2}\sum_{k=1}^{K_2}(f_j^{[k]}(z))^2 - (\hat{f}_j(z))^2\right)^{1/2}$$

where $f_j^{[k]}(z)$ is the estimate of $f_j(z)$ based on the coefficients $E[\beta|\gamma^{[k]},\lambda=\hat{\lambda},y]$. Estimates for the 95% confidence intervals can therefore be derived as $\hat{f}_j(z) \pm 1.96\hat{s}_j(z)$.

5 Empirical Results

We estimated the regression model for the data arising from each of the six monitoring sites. Starting at a variety of initial states, the Markov chain appears to converge reliably for each of the six data sets. For example, figure 3 contains some summaries of the Markov chain iterates resulting from the estimation with the Aldine site data, with initial state $\gamma = (0,0,\ldots,0)', \lambda = 0$. Figure 3 (a) demonstrates that the posterior probability $p(\gamma^{[j]}|y)$ converges to a stable distribution. Figures 3 (b) and (c) provide plots of the number of non-zero coefficients $q_{\gamma^{[j]}}$ and the cardinality of $\mathcal{S}$, respectively. These plots highlight that, at any particular iteration, there are around 25-40 non-zero regression coefficients and that the sampler only focuses on around 140-190 terms. Therefore, the sampler undertakes about one thirteenth of the number of generations required by a Gibbs sampler that generates all the binary indicators.

—Figure 3 About Here.—

Table 4 contains a summary of the normalized data transformation estimates for the ozone readings collected at all six monitoring sites. Four are logarithmic transformations, while the other two are very similar, with $\lambda = 1/4$ for the Dallas North data and $\lambda = -1/4$ for the Northwest Houston data. These confirm the type of transformations imposed on such data by several previous studies.

—Table 4 About Here.—

5.1 Meteorological Effects

Figure 4 plots surface slices of the estimate of the trivariate surface f_1 arising from the Aldine data. The slices are in humidity, with the nine panels corresponding to the $n/10, 2n/10, \ldots, 9n/10$th value of the sorted humidity variable. The surface slices all show an interaction between TR and WS, in that it is a combination of high temperature range and low wind speed that results in high ozone levels. There is also a strong humidity effect, with high overall ozone readings when the humidity level is low. Moreover, the effect of humidity appears to occur as an interaction with the temperature range and wind speed, with the surface slices altering substantially as humidity levels increase. For example, at the higher humidity levels the temperature range is low and does not have much impact on ozone formation, while high wind speeds tend to disperse the precursors to ozone formation very effectively at high humidity levels. At the lower humidity levels, the temperature range has a large effect on ozone formation, though even in combination with high wind speeds the precursors to ozone formation are not dispersed as effectively as at high humidity levels.

—Figures 4 and 5 About Here—

Figure 5 provides a plots surface slices of the function $1.96\hat{s}_1(TR, WS, HMD)$ arising from the estimation with the Aldine data. To enable a comparison, the scale on the vertical axis is the same as that for the function estimate $\hat{f}_1$ found in figure 4. This indicates that the trivariate confidence intervals are quite tight, although the standard error values are higher in areas of the domain where the data are more sparse and at the boundary of the convex hull of the data. The latter is the same boundary value effect frequently seen in bivariate and univariate function estimation.

The estimates of the trivariate surfaces are remarkably consistent across sites, even though some are located far apart, suggesting this is a fundamental meteorological determinant of ozone formation. For example, figure 6 plots the corresponding surface slices arising from the estimates at the Fort Worth Keller monitoring site. Note that for figures 4 and 6, the domain of the estimates will differ because we have plotted the function estimates over the convex hull of the each data set. These two trivariate surface estimates reveal the same basic non-linear interaction in the three variables– a profile that is also confirmed by the estimates from the remaining four sites.

——Figure 6 About Here.——

Table 5 provides the ranges of the estimates of all the functions at each of the sites and demonstrates that, by this measure, humidity, wind speed and temperature range have the greatest impact on ozone levels. The table demonstrates that the wind direction variables corresponding to $\hat{f}_2$, $\hat{f}_3$ and $\hat{f}_4$ are relatively minor in comparison to the meteorological effects captured by $\hat{f}_1$. We tried replacing the wind direction variables with their interaction with wind speed (that is, use the variables $WD_i * WS$) but this did not affect the results in a noticeable manner.

——Table 5 About Here.——

Figure 7 plots the estimates of these functions for the Aldine data, along with the approximate 95% confidence intervals. Notice that the intervals are quite tight (reflecting the fairly large sample size used) and that they are tighter for lower values of WD_i. This is due to the non-uniform distribution of the wind direction variables, with observations clustered at, or close to, $WD_i = 0$. As the wind increasingly blows from the South/West and North/East (that is, higher values of WD_2 and/or WD_3) there is a decrease in ozone formation. This may be because cleaner air is being blown in, compared to the East/South quadrant where the precursors to ozone are thought to be blown in from the Beaumont shipping channel. However, as the wind increasingly blows from the West/North there is a more indeterminant effect.

—–Figure 7 About Here—–

Notice that $\hat{f}_3$ and $\hat{f}_4$ are distinctly nonlinear, while the estimate $\hat{f}_1$ is close to linear. Table 6 summarizes the wind direction estimates by outlining whether they were noticeably nonlinear and the nature of the function. The summary is coded as AB, where A represents linear (L) or nonlinear (N) and B represents increasing ($\uparrow$) decreasing ($\downarrow$) or indeterminant ($\rightarrow$) levels of ozone as WD_i increases. They reveal that the effects are often nonlinear, which is not surprising as wind blowing constantly from one direction could result in any locally formed precursors to ozone being blown clear of the monitoring site, while wind blowing only partially from any single direction may not result in the precursors being blown clear of the monitoring site. In short, the relationship between these wind direction variables and ozone levels are probably smooth, but potentially prone to nonlinearity.

—–Table 6 About Here—–

5.2 Seasonal and Trend Estimates

Figure 8 plots the seasonal estimate $\hat{f}_5$ for all six sites, along with the respective 95% confidence intervals. Five of the six sites have the same basic profile, with a decrease in residual ozone levels during May and June, stable levels in July, August and September and a further decrease in October. Only the data from the Dallas North monitoring site has a noticeably different profile, with stable or increasing residual ozone levels in May and June, followed by a gradual decrease until October. The high degree of similarity in the profiles of $\hat{f}_5$ at each of the sites could result from meteorological variation not captured by $f_1, \ldots, f_4$ and not random over the period May-October.

——Figures 8 and 9 Here.——

The trend estimates are plotted in figure 9 and reveal substantial variation in trends at the six monitoring sites. This can be partly explained by the different environments in which the stations are located. For example, the estimate in the trend at the Beaumont monitoring site is highly variable, which could be due to the concentration of a sizable part of the world's petrochemical plant in the area. Such industrial activity is thought to have a high impact on the formation of ozone and the level of industrial activity is not even throughout the period 1980-1997. In particular, the years 1991-1995 appear to be periods in which ozone levels were high at all sites, apart from Clinton. This could well be due to increased economic activity during this period, relative to the previous period. Lastly, figure 10 provides plots of the annual means of the raw ozone at each site. A comparison with the trend estimates in figure 9 reveals that the meteorological variation appears to mask the undertlying trend values quite substantially.

——Figure 10 About Here.——

6 Comparison with MARS

A popular alternative method for estimating such multivariate nonparametric regression models is MARS (multivariate adaptive regression splines) proposed by Friedman (1991). Other methods include that of Stone, Hansen, Kooperburg and Truong (1996) which is similar to MARS in the regression case. We compare the Bayesian estimate to that obtained using MARS via a simulation based on the design of the data from the Clinton monitoring site. We chose this dataset simply because it has the smallest sample size and therefore is the fastest on which to run the simulation.

In this simulation we estimated the regression model at (3.1) using MARS (version 3.6). The procedure does not allow for estimation of a data transformation and therefore we simply used the transformed data as the dependent variable. The MARS program uses somewhat different bases for the various components, including a tensor product regression spline basis for the trivariate function f_1, univariate regression splines for the additive functions f_2, f_3 and f_4 and dummy variable bases for f_5 and f_6. However, linear regression spline terms are used instead of the cubic regression spline terms during the search for suitable knot locations to provide speedy computations. Once these locations have been determined, estimation is undertaken with cubic regression spline terms using these knot locations.

To compare both the Bayesian estimate (BAYES MODEL) and that provided by MARS (MARS MODEL), we simulated data from both and fit both estimators to the data. Therefore, the simulated datasets have the same design as the original Clinton data (that is, the same independent variable values), though we simulated fifty replicates of dependent variables form each of the two models. We measured the performance of the estimators on reproducing both true models by calculating the following distance measure for each replicated dataset and both estimators

$$ISE = \frac{1}{1804} \sum_{i=1}^{1804} (\tilde{y}_i - \hat{y}_i)^2$$

Here, $\tilde{y}_1, \ldots, \tilde{y}_{1804}$ are the fitted values from the original data (obtained using either the BAYES MODEL or MARS MODEL), while $\hat{y}_1, \ldots, \hat{y}_{1804}$ are the fitted values obtained from the simulated data using either procedure. Lower values of this distance measure indicate that the estimate is closer to the true model.

——Figure 11 About Here.——

Figure 11 provides the log(ISE) values for both the estimators for data generated from the BAYES MODEL and MARS MODEL. It can be seen that the Bayesian estimator more faithfully reproduces the BAYES MODEL, which is to be expected as this model is the Bayesian estimate from the original data. However, the Bayesian estimator also more accurately estimates the MARS MODEL. This is remarkable as the Bayesian procedure is not only a different estimation procedure, but is using a different basis than the MARS MODEL. These results suggest that the Bayesian approach is substantially more efficient than MARS when applied to the regression model at (3.1) with the design presented by the Clinton data. This corresponds to the simulation results presented in Smith and Kohn (1997) for the case of bivariate surface estimation.

7 Conclusion

This paper has a number of objectives. First, it demonstrates that the proposed Bayesian nonparametric regression method can be applied to a complex regression problem in a larger sample size environment. This methodology is very general, with the user able to select which bases with which to work. In particular, we demonstrate this by using a trivariate radial basis to model the response of the key meteorological variables.

Second, a high dimensional basis representation is obtained by locating the basis terms at each of the design points. Such a large basis cannot be handled effectively using the 'one at a time' Gibbs sampling scheme discussed in Smith and Kohn (1996) and requires an alternative, such as the focused sampling approach. Third, by using a fractional prior for β_γ at A1, the methodology is fully data-driven. Moreover, an appropriate transformation of the dependent variable from a discrete set of location and scale invariant candidate transformations is estimated along with the regression surfaces.

Fourth, we have contributed an empirical study that improves understanding of the determinants of tropospheric ozone levels. Here, we have estimated the functional form of the dependence of daily ozone maxima on key meteorological variables in a more general and flexible way than previous authors who assume a parametric or additive structure. These estimates are made more meaningful by their consistency across six different monitoring sites in areas of concern in Texas. In addition, we provide estimates of meteorologically adjusted long-term trends and show they differ substantially from that observed in the unadjusted ozone data.

Lastly, we compare our estimator to MARS, one of the few alternative nonparametric regression procedures capable of estimating such a multivariate regression model with the sample sizes in our data. We do this using a simulation study based on the design of the data at one of the monitoring sites. The results indicate that for this problem, contemporary Bayesian nonparametric regression is substantially more reliable than MARS.

References

Bloomfield, P, Royle, A. and Yang, Q., (1993), "Accounting for Meteorological Effects in Measuring Urban Ozone Levels and Trends", National Institue of Statistical Sciences, Technical Report Number 1.

Box, G., and Cox, D., (1982), "An analysis of transformations revisited, rebutted", *Journal of the American Statistical Association*, vol. 77, 209-210

Carroll, R., Chen, R., George, E., Li, T., Newton, H., Schmiediche, H., and Wang, N., (1997), "Ozone Exposure and Population Density in Harris County, Texas", *Journal of the American Statistical Association*, vol. 92, no. 438, 392-415

Cleveland, W., E. Grosse, and Shyu, W., (1992), "Local regression models", in Chambers, J and Hastie, T., eds., Statistical Models in S, Pacific Grove, California: Wadsworth, pp. 309-373.

Cox, W., and Chu, S, (1992), "Meteorologically Adjusted Ozone Trends in Urban Area: A Probability Approach", *Atmospheric Environment*, 27B, 425-434

Denison, D., Mallick, B. and Smith, A., (1998), *Journal of the Royal Statistical Society*, Series B, vol. 80, 331-350.

Friedman, J., (1991), "Multivariate Adaptive Regression Splines", *The Annals of Statistics*, vol. 19, no. 1, 1-67

Gu, C., Bates D., Chen, Z., and Wahba, G., (1989), "The computation of GCV functions through Householder tridiagonalization with application to the fitting of interaction spline models," *SIAM J. Matrix Analysis and Applications*, 10, 457-480

Green, P., (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination", *Biometrika*, 82, 711-732

Holmes, C. and Mallick, B., (1997), "Bayesian Radial Basis Functions of Unknown Dimension", preprint

Luo, Z. and Wahba, G., (1997), “Hybrid Adaptive Splines”, *Journal of the American Statistical Association*, vol. 92, no. 437, 107-116

O’Hagan, A., (1995), “Fractional Bayes factors for model comparison” (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 57, 99-138

Powell, M., (1987), “Radial basis functions for multivariate interpolation: a review”, in Mason, J., and Cox, M., (eds.), *Algorithms of Approximation*, Oxford, Clarendon Press, 143-167

Niu, X., (1996), “Nonlinear Additive Models for Environmental Time Series, With Applications to Ground-Level Ozone Data Analysis”, *Journal of the American Statistical Association*, vol. 91, no. 435, 1310-1321

Nychka, D., Yang, Q. and Royle, J.A., (1998), “Constructing spatial designs using regression subset selection”, preprint.

Schwarz, G., (1978), “Estimating the Dimension of a Model”, *The Annals of Statistics*, vol. 6, no. 2, 461-464

Shively, T., (1990), “An analysis of long-term trend in ozone levels at two Houston, Texas sites”, *Atmospheric Environment*, 24B, 293-301

Shively, T. and Sager, T., (1997), “A semiparametric regression approach to adjusting for meteorological variables in pollution trends”, working paper

Smith, R. and Huang, L., (1993) “Modelling High Threshold Exceedances of Urban Ozone”, National Institute of Stastical Sciences, Technical Report Number 6.

Smith, M., and Kohn R., (1996), “Nonparametric regression via Bayesian variable selection,” *Journal of Econometrics*, vol. 75, no. 2, 317-344

Smith, M. and Kohn, R., (1997), “A Bayesian Approach to Nonparametric Bivariate Regression”, *Journal of the American Statistical Association*, vol. 92, no. 440, 1522-1535

Stone, C., Hansen, M., Kooperburg, C., Truong, Y., (1996), "Polynomial splines and their tensor products in extended linear modeling", *Annals of Statistics*, 25, 1371-1470

Tierney, L., (1994), "Markov chains for exploring posterior distributions", *The Annals of Statistics*, vol. 22, 1701-1762

Wahba, G., (1990), *Spline models for observational data*, Philadelphia, SIAM

Wong, F., Hansen, M., Kohn, R., Smith, M., (1997), "Focused sampling and its application to nonparametric regression", preprint.

	Quadrants			
Monitoring Site	Quad 1 (WD_1)	Quad 2 (WD_2)	Quad 3 (WD_3)	Quad 4 (WD_4)
Aldine	90-180	180-270	270-0	0-90
Clinton	325-45	45-112.5	112.5-202.5	202.5-325
Northwest Houston	90-180	180-270	270-0	0-90
Dallas North	135-225	225-315	315-45	45-135
Forth Worth Keller	68-158	158-248	248-338	338-68
Beaumont	90-180	180-270	270-0	0-90

Table 1: Definitions of the quadrant directions in degrees, where 0=North and 180=South.

Monitoring Site	Period Data Collected	Sample Size (n)	% missing
Clinton	1/5/83-31/10/95	1804	18.30%
Aldine	1/5/80-31/10/97	2614	21.07%
Beaumont	25/9/80-30/9/97	2373	24.28%
Dallas North	1/5/80-31/10/97	2737	17.36%
Fort Worth Keller	1/5/83-31/10/97	2309	16.34%
Northwest Houston	1/5/81-31/10/97	2433	22.22%

Table 2: Period over which data was collected, resulting number of full observation and percentage of data missing within the respective collection periods for all six sites.

Domain		Clinton	Aldine	Beaumont	Dallas N.	Ft. Worth Keller	NW. Houston
TR, WS, HMD	p_1	1476	2247	2077	2522	2152	2230
WD_2	p_2	40	36	36	31	41	48
WD_3	p_3	47	35	33	29	31	33
WD_4	p_4	43	37	34	28	31	44
MN	p_5	6	6	6	6	6	6
YR	p_6	13	18	18	18	15	17
Total	p	1625	2379	2204	2634	2276	2378

Table 3: Total number of basis terms p_i for each component function f_i for the data corresponding to each of the six sites. Here, p is the total number of basis terms, including the global intercept α.

Monitoring Site	$\hat{\lambda}$	$t_{\hat{\lambda}}$	$a_{\hat{\lambda}}$	$b_{\hat{\lambda}}$
Clinton	0	$\log(OZ+1)$	-5.0476	6.1658
Aldine	0	$\log(OZ+1)$	-8.0035	7.7103
Beaumont	0	$\log(OZ+1)$	-4.3399	5.7708
Dallas North	0.25	$(OZ+1)^{1/4}$	-20.8463	17.1196
Fort Worth Keller	0	$\log(OZ+1)$	-7.3976	7.3989
Northwest Houston	-0.25	$(OZ+1)^{-1/4}$	33.9083	43.7684

Table 4: Estimated transformations for the data at all six sites.

	Range of Function Estimate					
Monitoring Site	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_4$	$\hat{f}_5$	$\hat{f}_6$
Aldine	15.154	2.8983	2.029	2.655	1.413	1.966
Clinton	10.189	1.5789	1.094	1.135	1.631	1.484
Beaumont	7.518	1.3882	1.125	0.965	1.314	5.074
Dallas North	16.142	0.703	1.119	2.188	1.518	1.683
Fort Worth Keller	9.936	1.656	2.462	2.013	1.505	1.258
Northwest Houston	13.596	3.457	1.583	1.790	1.491	2.681

Table 5: Range of the estimated functions $\hat{f}_1, \ldots, \hat{f}_6$ obtained from the data collected at all six monitoring sites.

Monitoring Site	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_4$
Aldine	L↓	N→	N↓
Clinton	N↑	N→	N→
Beaumont	L↓	N→	N→
Dallas North	L↑	N↓	L↓
Fort Worth Keller	L↓	L↓	L↓
Northwest Houston	L↓	N→	N↓

Table 6: Summary of the estimates for $\hat{f}_2$, $\hat{f}_3$ and $\hat{f}_4$ for the data collected at each of the six data sites. The profile of these functions is summarised as a pair AB, where A represents linear (L) or nonlinear (N) and B represents increasing (↑) decreasing (↓) or indeterminant (→) levels of ozone as WD_2, WD_3 and WD_4 increases.

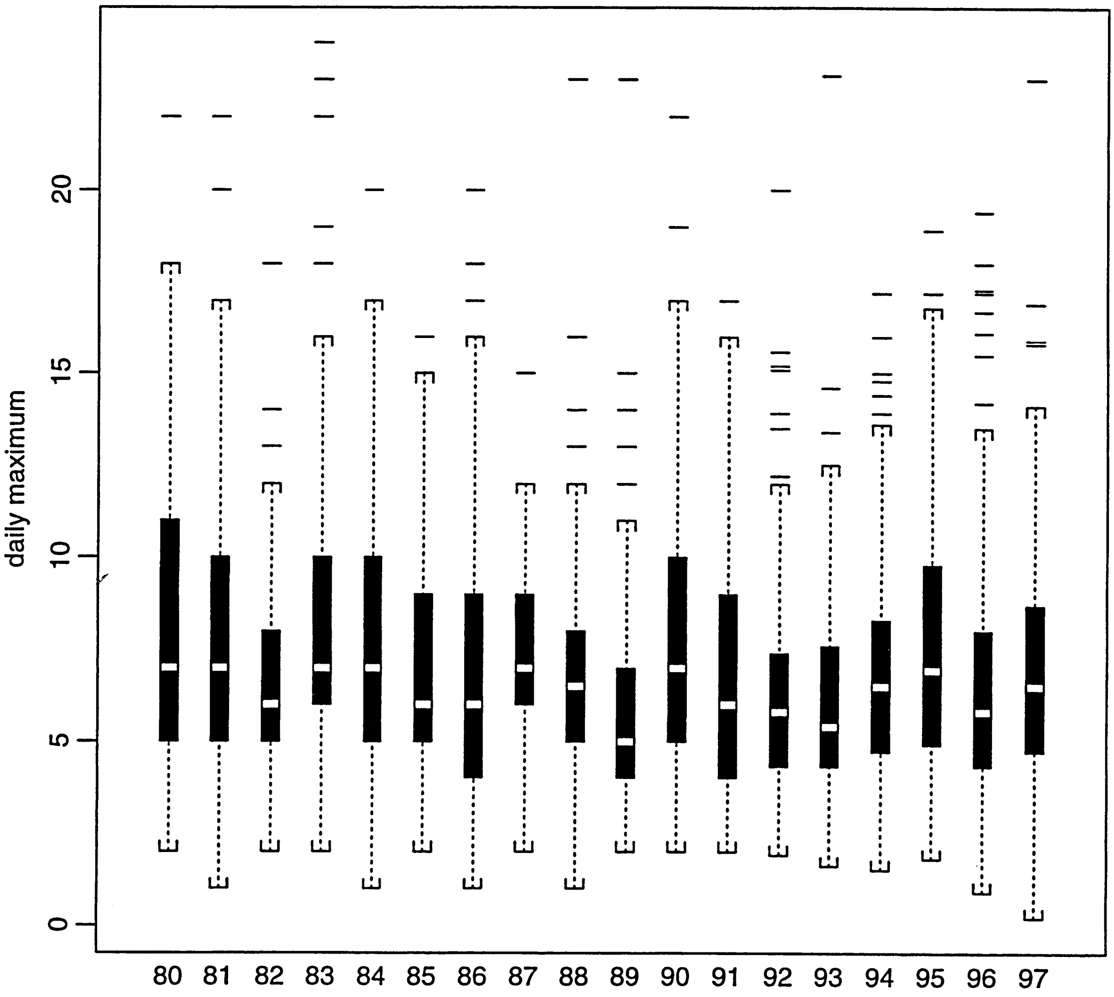

Figure 1: Boxplots of the daily maxima of hourly tropospheric ozone concentrations (in parts per hundred million) during the period 1980-1997 at the Aldine monitoring site. Note that the US Environmental Protection Agency's national ambient air quality standard in 12 pphm.

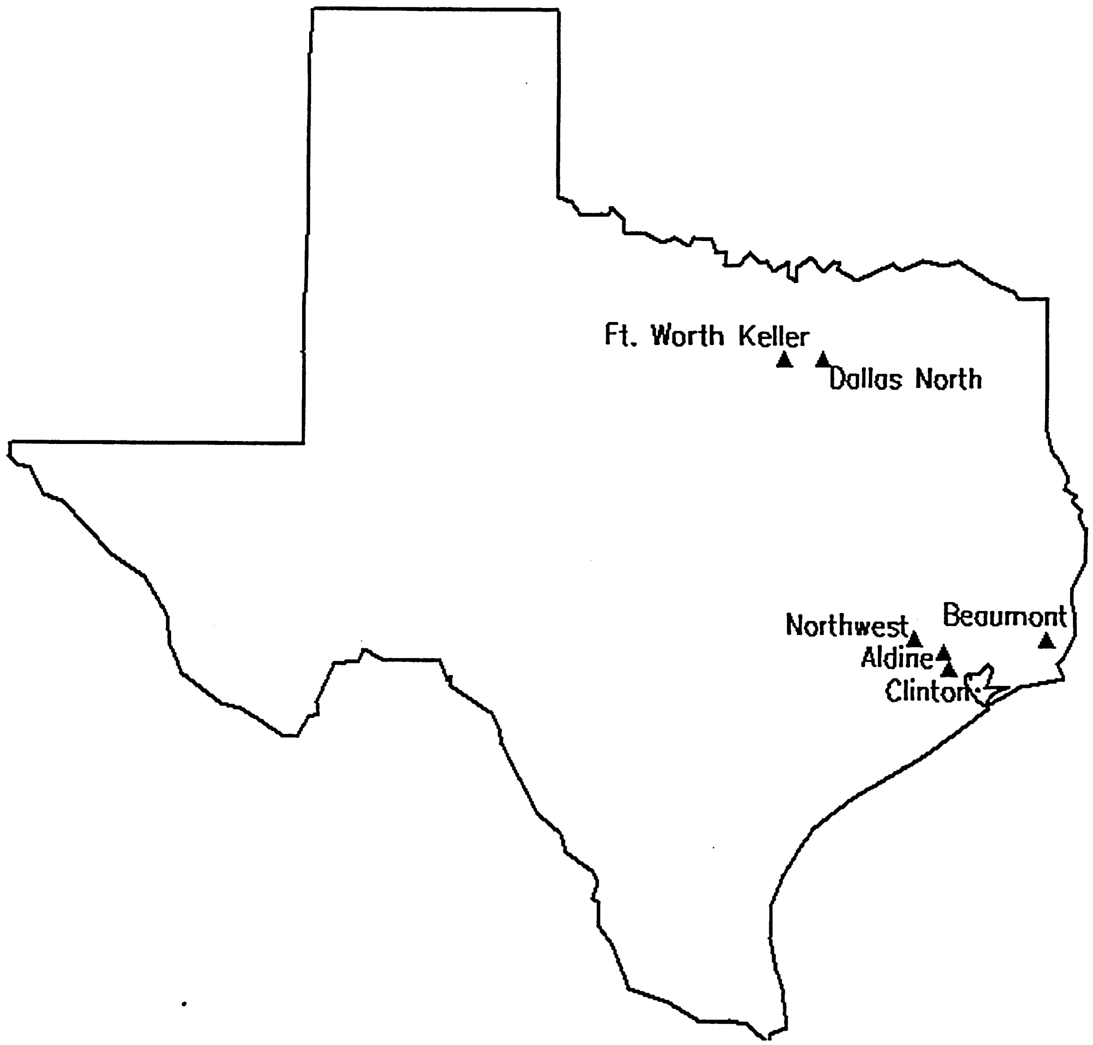

Figure 2: Map providing location of monitoring sites. All six are located in areas of particular concern to the Texas Natural Resource Conservation Commission.

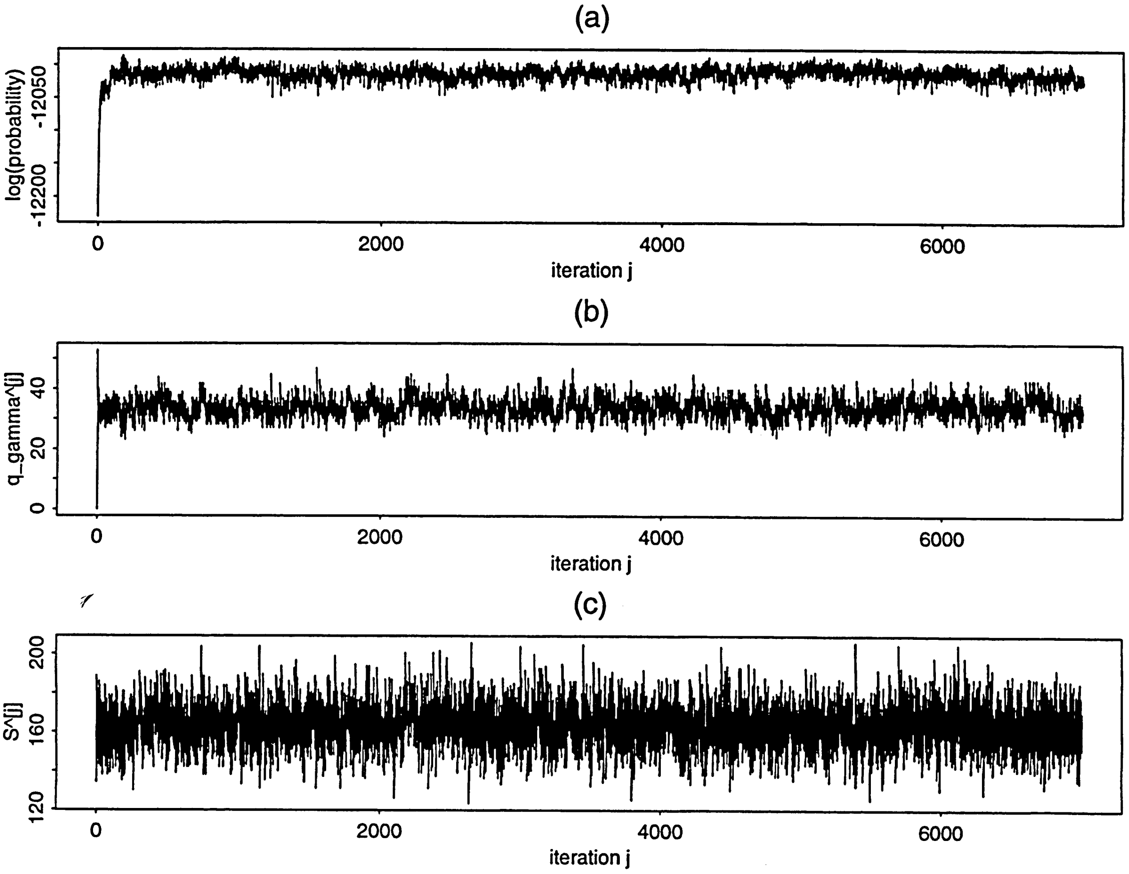

Figure 3: Summaries of the Markov chain iterates for the estimation of the regression model with the Aldine monitoring site data. (a) The posterior probability $p(\gamma^{[j]}|y)$. (b) The number of non-zero coefficients $q_{\gamma^{[j]}}$. (c) The cardinality of the focus set $\mathcal{S}$. The plots are produced for iterates from the warmup and two sampling periods.

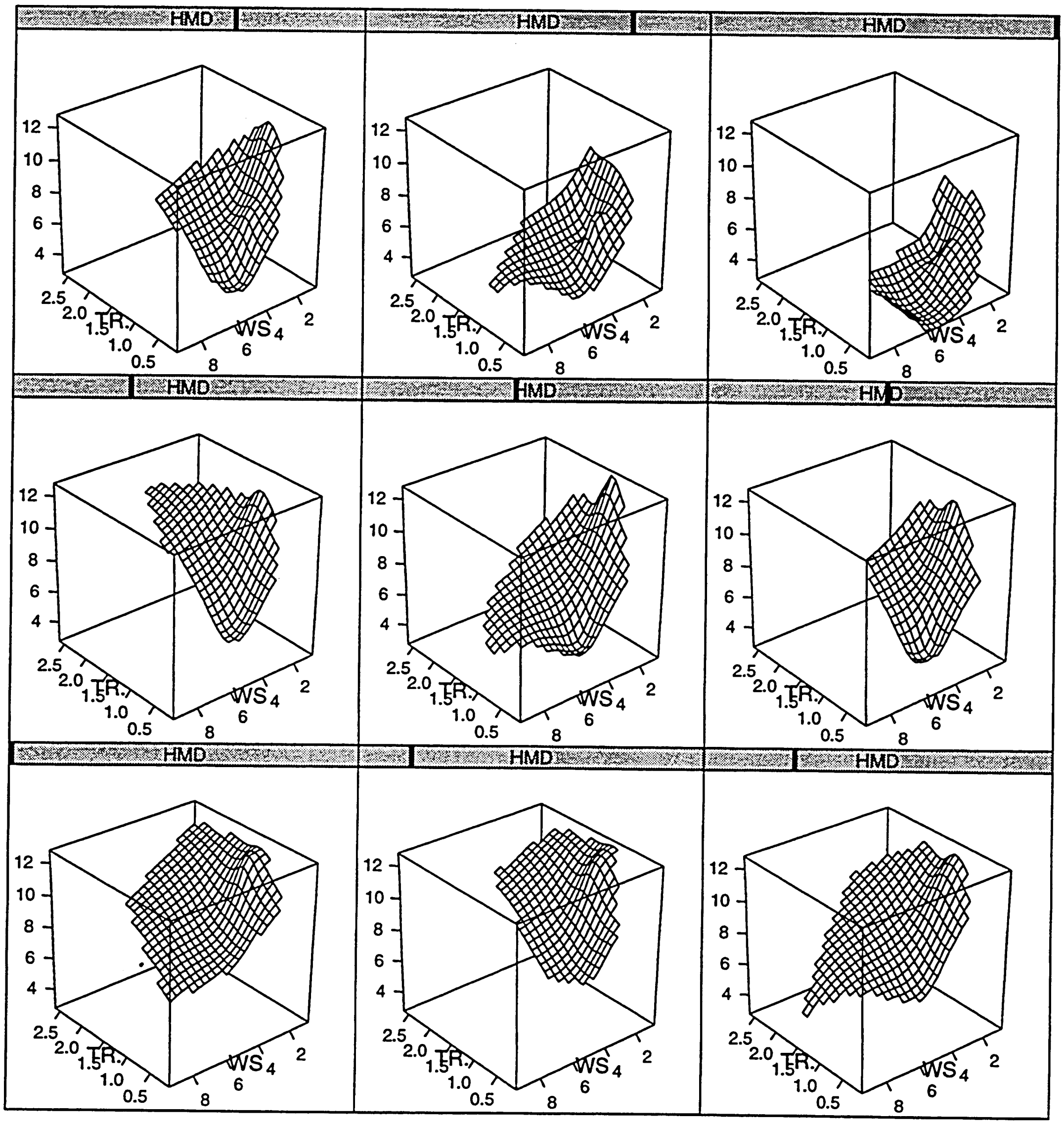

Figure 4: The trivariate surface estimate $\hat{f}_1(TR, WS, HMD)$ estimated from the Aldine site data. The nine surfaces plotted are of $\hat{f}_1(TR, WS, HMD = x)$ for $x = n/10, 2n/10, \ldots, 9n/10$th ordered value of HMD. The bottom left hand panel corresponds to low humidity, while the top right hand panel corresponds to high humidity.

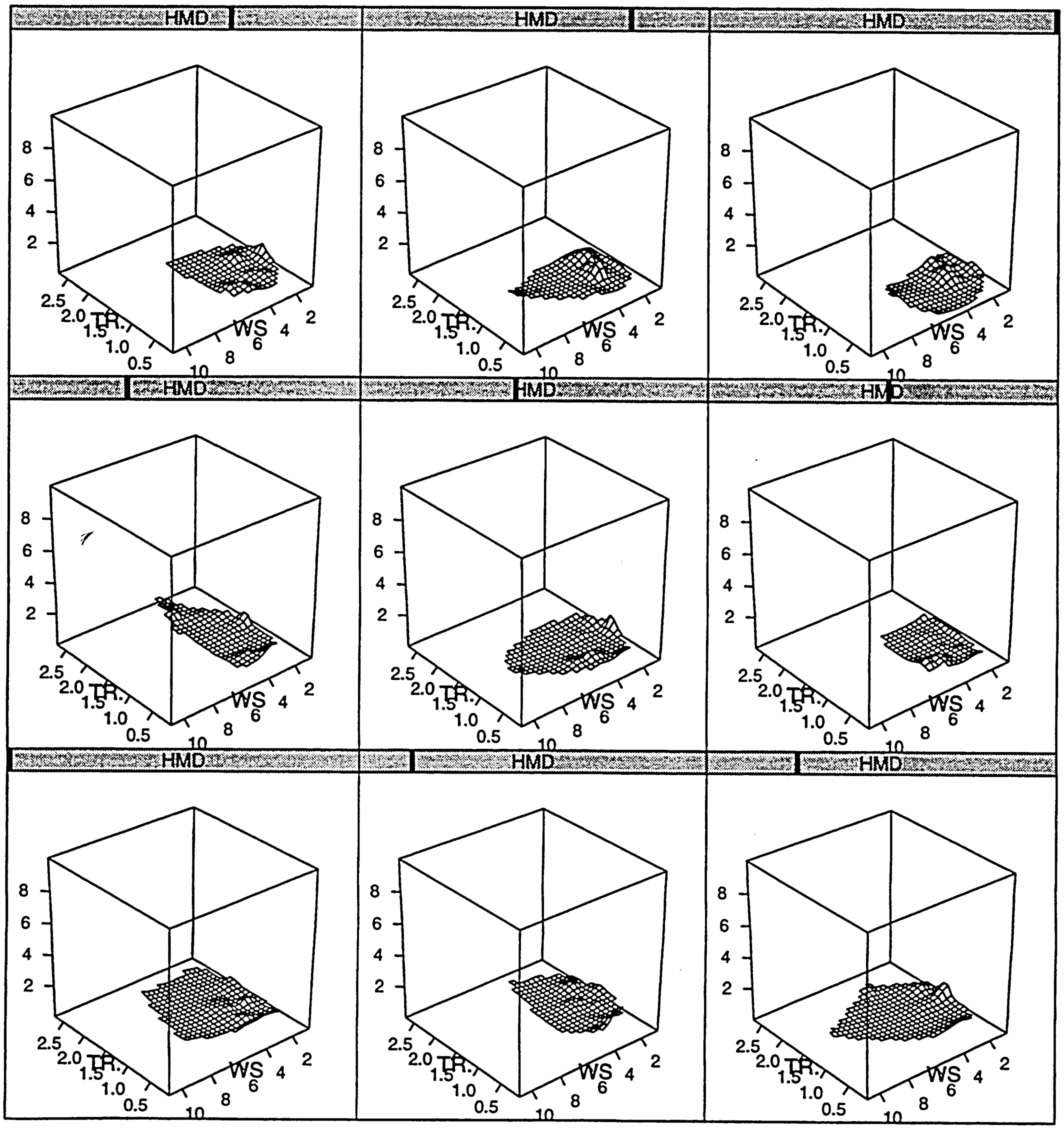

Figure 5: The estimate of the upper confidence interval $1.96\hat{s}_1$ for the trivariate function f_1 estimated from the data collected at the Aldine monitoring site. The nine surfaces plotted are of $1.96\hat{s}_1(TR, WS, HMD = x)$ for $x = n/10, 2n/10, \ldots, 9n/10$th ordered value of HMD. The range of the vertical axis is set to that used in the plot of $\hat{f}_1$ in figure 4 to enable comparison.

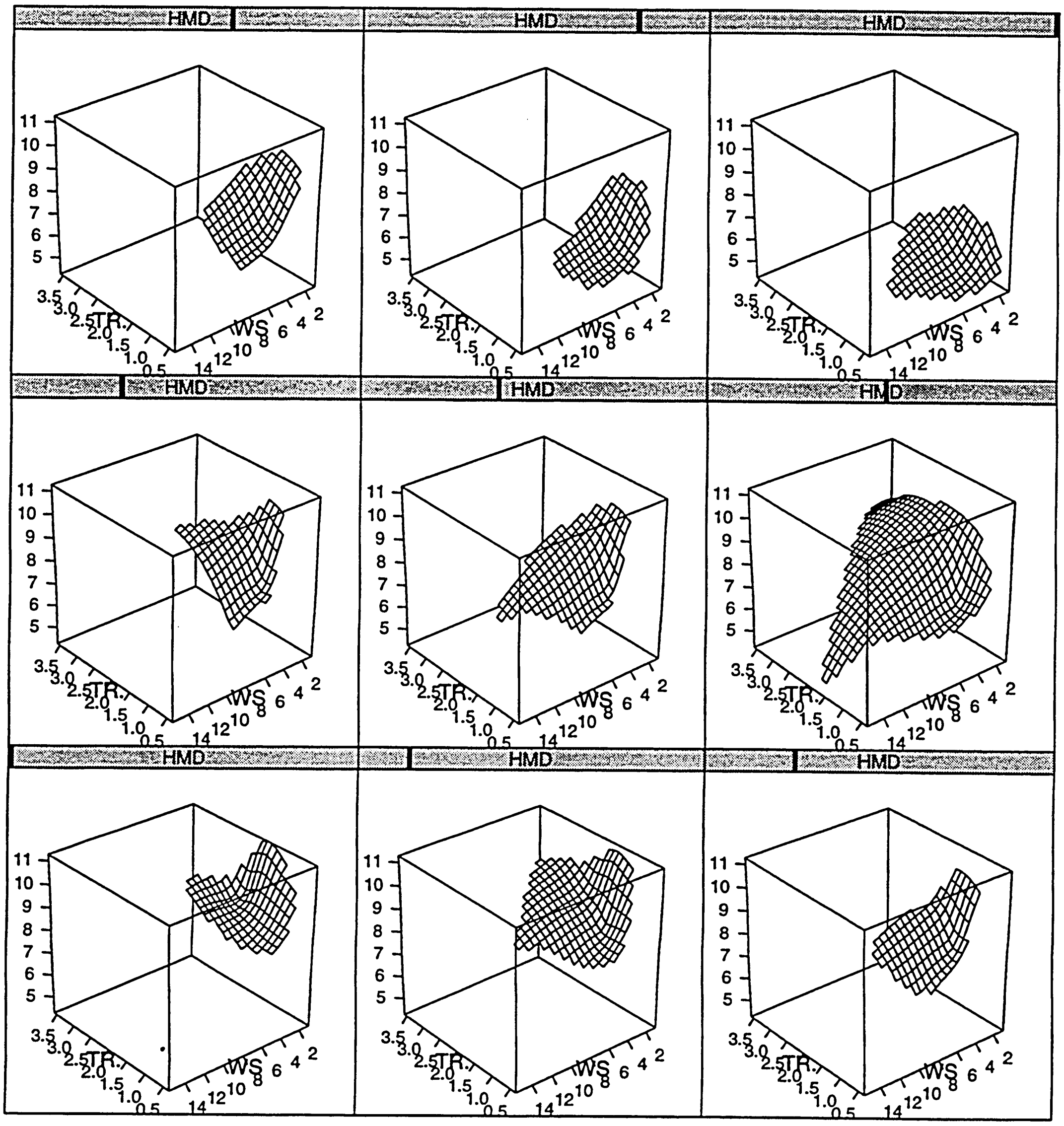

Figure 6: The trivariate surface estimate $\hat{f}_1(TR, WS, HMD)$ estimated from the Forth Worth Keller site data. The nine surfaces plotted are of $\hat{f}_1(TR, WS, HMD = x)$ for $x = n/10, 2n/10, \ldots, 9n/10$th ordered value of HMD. Note that the humidity levels in which the slices are made will therefore differ slightly from that found in the surface slices from the Aldine site.

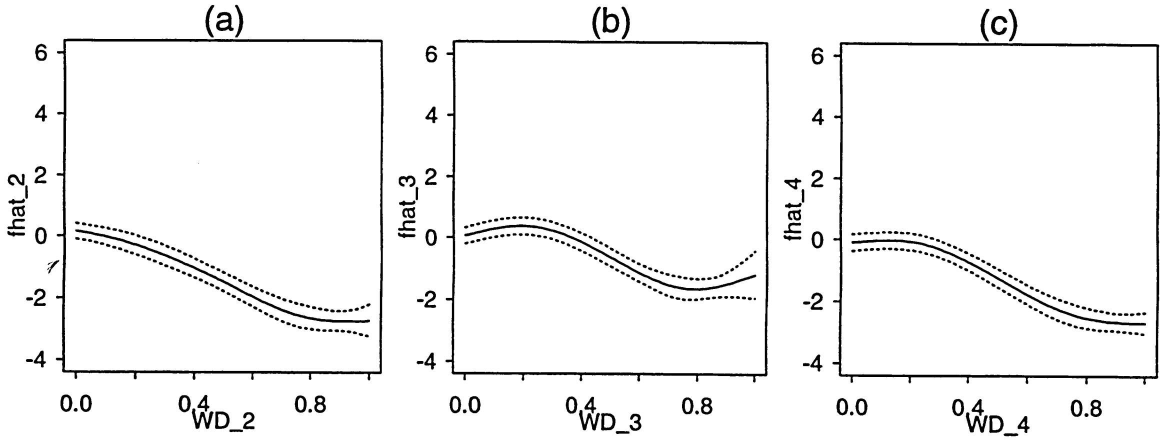

Figure 7: (a)-(c) Plots of $\hat{f}_2, \hat{f}_3$ and $\hat{f}_4$ (bold lines), respectively, for the data collected at the Aldine monitoring site. Also plotted are the 95% confidence intervals (dotted lines).

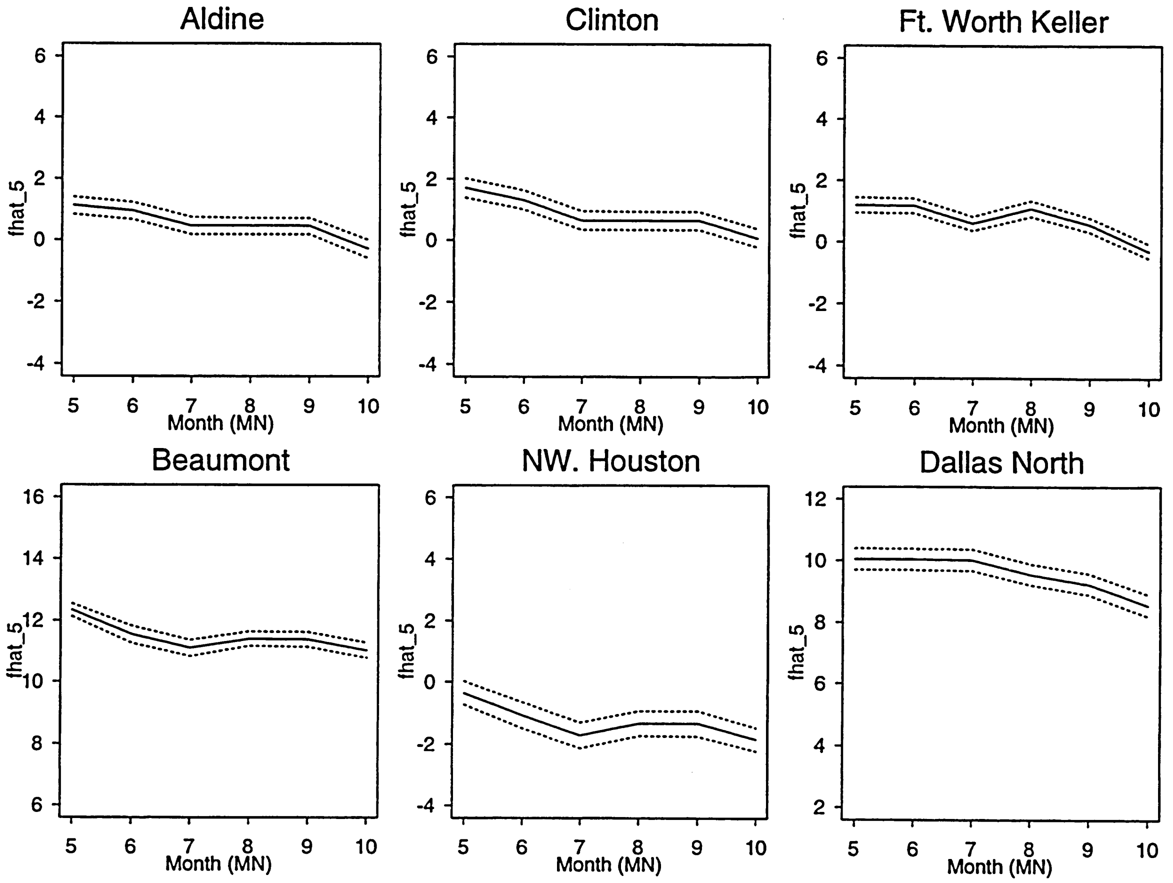

Figure 8: The estimated residual seasonal effect for each of the six monitoring sites. In each panel, the bold line is the estimate of $\hat{f}_5$, while the dotted lines provide 95% confidence intervals.

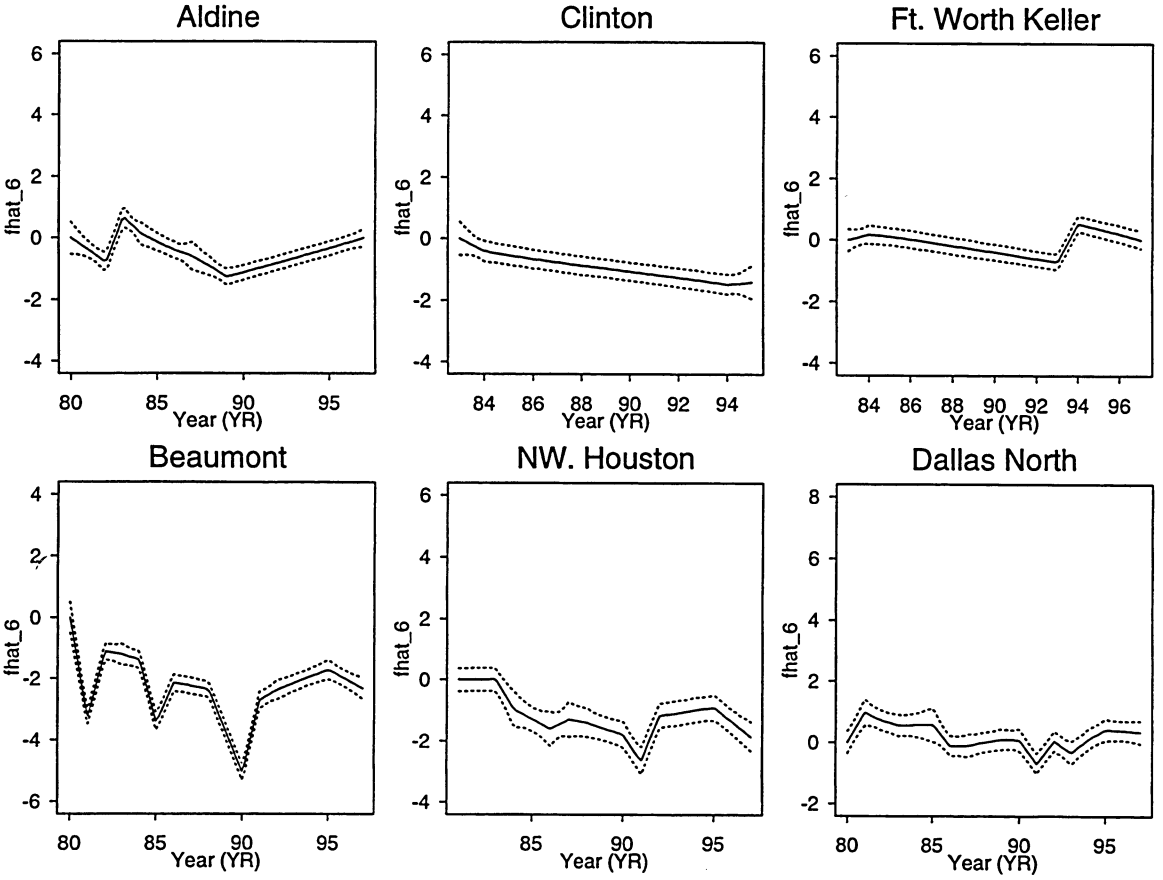

Figure 9: The estimated long-term trend in ozone levels for each of the six monitoring sites. In each panel, the bold line is the estimate of $\hat{f}_6$, while the dotted lines provide 95% confidence intervals.

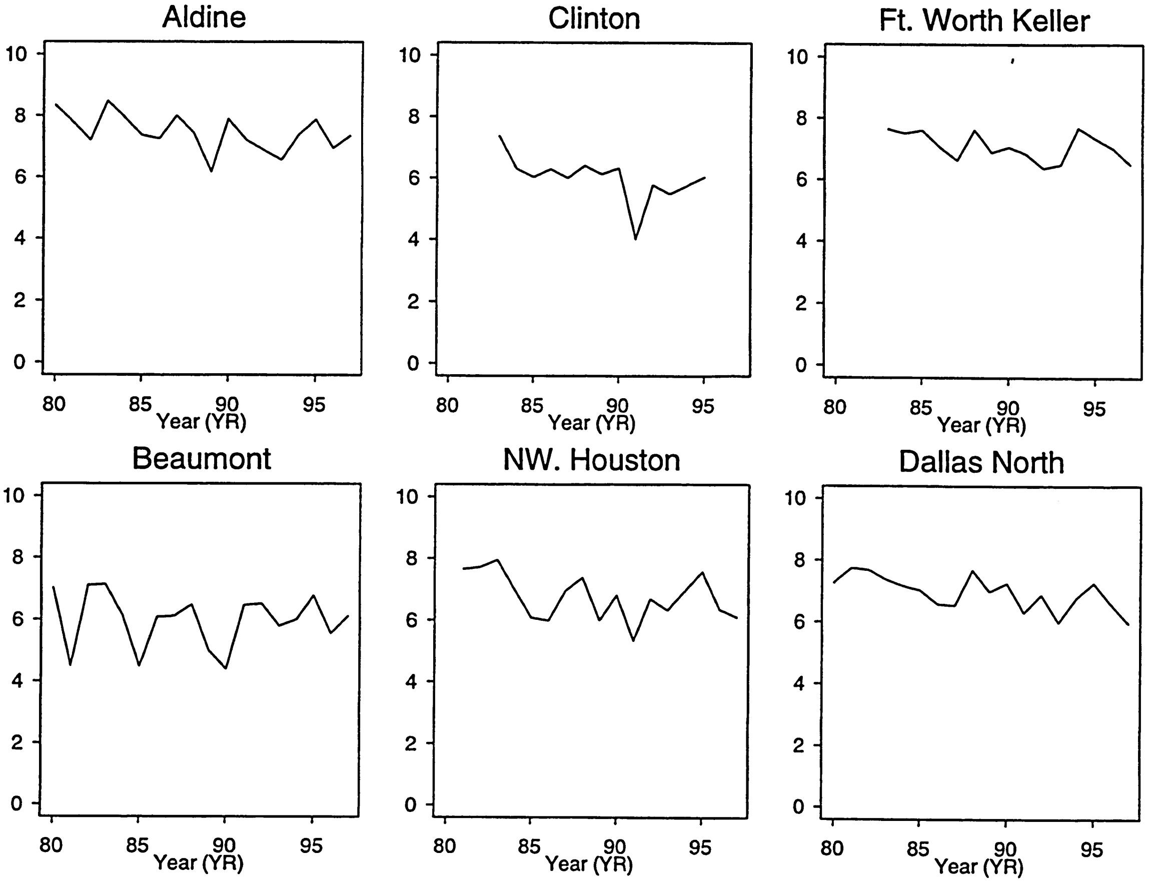

Figure 10: Plots of the mean values of the transformated ozone data collected at each of the sites. The transformations used are the estimated data transformations and enable a comparison with the trends plotted in figure 9.

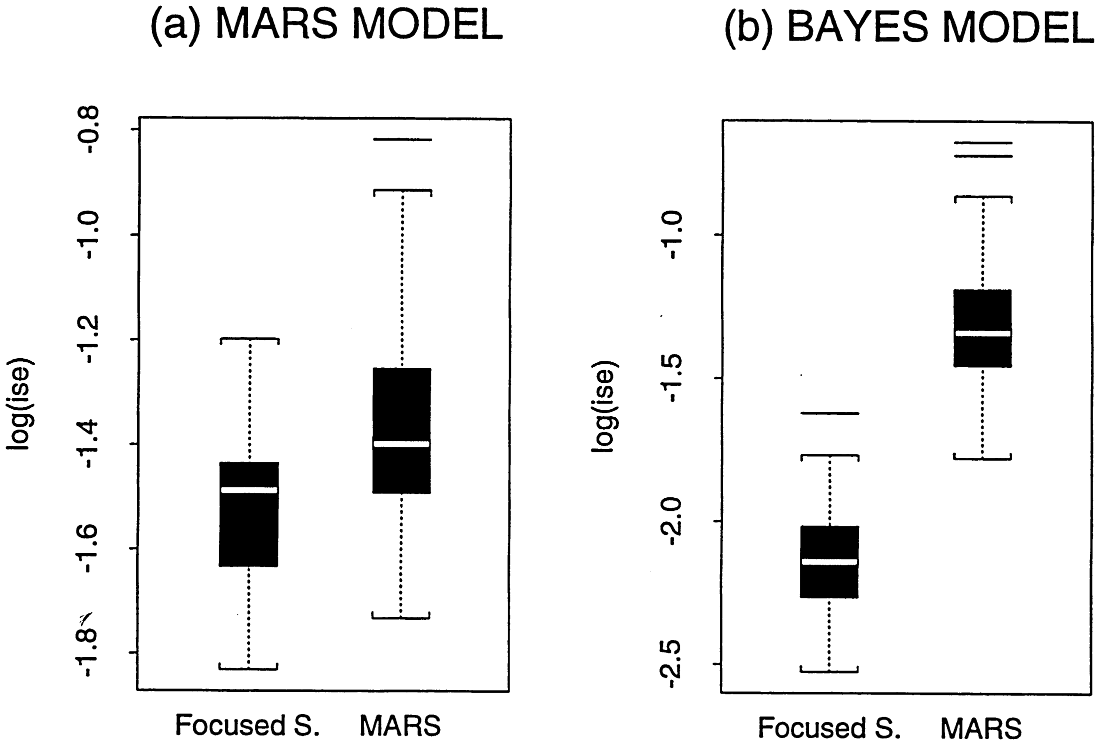

Figure 11: Simulation comparison of the MARS and focused sampling based estimates. Panel (a) is for data generated from the MARS MODEL and panel (b) is for data generated from the BAYES MODEL. Each boxplot is constructed from the fifty log(ISE) values resulting from the fifty simulation replicates for each of the two models.