BAYESIAN MULTIVARIATE SPATIAL INTERPOLATION: APPLICATION AND ASSESSMENT

by

Weimin Sun, Nhu D. Le,
James V. Zidek & Rick Burnett

Technical Report #146

March 1995

WS/PR:ddp

Bayesian Multivariate Spatial Interpolation: Application and Assessment

Weimin Sun[1], Nhu D. Le[1,2], James V. Zidek[1] and Rick Burnett[3]
[1] Department of statistics
University of British Columbia
[2] Biometry Section
British Columbia Cancer Agency
[3] Health Canada

March, 1995

Abstract

This paper implements and empirically assesses a Bayesian methodology for the spatial interpolation of random multivariate Gaussian fields with unspecified covariance structure. One special feature of the method is that it does not require all sites to monitor the same set of pollutants. This feature is particularly relevant in environmental health studies where pollution data are often pooled together from several monitoring networks which may or may not monitor the same set of pollutants. The methodology is applied to the data in the Province of Ontario, where monthly average concentrations for summer months of nitrogen dioxide (NO_2), ozone (O_3) , sulphur dioxide (SO_2) and sulfate (SO_4) are available for the period from January 1 of 1983 to December 31 of 1988 at 31 ambient monitoring sites. Detailed descriptions of spatial interpolation for air pollutant concentrations at 37 approximate centroids of Public Health Units in the Province of Ontario using all available data are presented. Empirical assessment of the methodology is done by a crossvalidation study where each of the 31 sites is successively removed and the remaining sites are used to predict its concentration levels. The methodology seems to perform well.

Keyword: Air Pollution; Bayesian Spatial Interpolation; EM Algorithm; Ozone; Sulphate; Nitrate; Sulphur Dioxide;

1 Introduction

Le, Sun and Zidek (1994, hereafter LSZ) recently proposed a Bayesian methodology for multivariate spatial interpolation. The method is particularly useful in environmental health

studies where vector-valued responses are only measured at designated sites at successive time points and data are generally not available at many locations of interest. One special feature of their approach is that it does not require all sites to monitor the same set of pollutants; that is, not all of a specific set of pollutants are observed at all sites. The missing pollutants at the monitoring sites are termed *missing-by-design*. Specifically they derive the multivariate predictive distribution for all pollutants at the non-monitored locations, along with the missing-by-design pollutants using all available data.

In this paper, we apply the LSZ methodology to Southern Ontario air pollution data. This kind of interpolation was needed in a recent study (not presented here) of the association between air pollution and respiratory morbidity in the population of Southern Ontario. Specifically, monthly average pollutant concentrations down to the level of a Public Health Unit (PHU) are needed. Since air pollution data from several monitoring networks are pooled together, not all sites monitor the same set of pollutants. Here it is assumed that the variation caused by different networks to the observations is negligible.

The pollutants under consideration included NO_2, SO_2, O_3 and SO_4. In general there are two kinds of air pollutants: (i) a primary pollutant, which is directly emitted by identifiable sources; (ii) a secondary pollutant, which is produced by chemical reactions within the atmosphere between pollutants and other constituents. SO_2, NO_2 are primary pollutants, O_3 and SO_4 are secondary pollutants. SO_2 is produced by burning of sulphur contained fuels and its level depends on the local emission sources, like burning fuel oil or smelting. NO_2 can be produced by high temperature combustion and so its level could depend on the local sources as well.

The secondary pollutants studied here are all produced by oxidation of primary pollutants. This oxidation is driven by ultra-violet radiation from sunlight and comprises chemical reactions that are temperature dependent. Since the chemical reactions proceed while the polluted air is being adverted by winds, secondary pollutants are generally more widespread than primary pollutants. We thus refer to secondary pollutants as regional. Since NO_2 can also be produced by oxidation, it could be considered as a regional pollutant. The dominant factor in producing NO_2 could be either local sources or oxidation depending on the atmospheric condition. The results from our analyses indicate that NO_2 behaves more like a primary pollutant in this case. Because of temperature dependence of the governing chemical reaction, O_3 levels are high in early afternoon and midsummer, low overnight and

in winter. The oxidation of SO_2 to SO_4 is dominated by photochemical processes in dry, warm atmospheres.

The proposed methodology is empirically assessed by cross-validation where the sites are successively removed one at a time and the remaining sites are used to impute the missing data. The results are fairly promising; some interesting findings are highlighted here and more details are given in other sections. As expected, the correlations between the observed and predicted values are higher for regional pollutants than for local pollutants. The correlations averaging over all sites for O_3 and SO_4 are .96 and .97 respectively; the correlations for NO_2 and SO_2 are .64 and .78 respectively. Figures 1 and 2 depict typical temporal patterns between the observed and predicted values for regional and local pollutants. For regional pollutants, the methodology performs very well in predicting both the concentration levels and the temporal patterns. This result is not too surprising since the regional pollutants tend to be highly spatially correlated. For local pollutants, there seem to be some biases in predicting the levels. This deficiency indicates the need for more data in the neighborhood when predicting the local pollutants. On the other hand, the temporal patterns seem to be tracked well even with limited amounts of data. This result is encouraging in that it may be possible to improve the current method to deal with the bias problem for local pollutants. For example, a prior distribution of the mean process where the levels at only neighborhood sites are used, could be adopted instead.

The proposed method allows for the use of all available data in its interpolation. One interesting question is whether anything can be gained from this appeoach in comparison with that of interpolating one pollutant at a time. Our results indicate that the current multivariate approach could provide substantial improvements over the univariate approach on regional pollutants. The relative reductions of the mean squared prediction errors from using the multivariate interpolation over the univariate interpolation are about 300% and 900% for O_3 and SO_4. This information gain is due to the high correlation between O_3 and SO_4 and hence observed values of one pollutant help to predict the other. It is interesting to note that the gain for SO_4 is substantially higher than that for O_3. This is so because there is more information available for O_3 in predicting SO_4 than vice-versa; 21 sites monitor O_3 compared to 10 for SO_4. The improvements are not as great for the local pollutants as expected. For example, the reduction for SO_4 is from 1.27 to 0.14 and for SO_2, from 0.76 to 0.62.

The paper is organized as follows. For completeness, we briefly describe the theory in Section 2. Section 3 describes its implementation on the data from southern Ontario. Then in Section 4, we look at how well it works. Section 5 looks at the gains we make from combining sites and pollutants in a single analysis.

2 Bayesian Interpolation Theory

Following LSZ, we assume a normal model for the conditional pollutant sampling distribution,

$$X \mid Z, B, \Sigma \quad \sim N_{sk \times n}(BZ, \Sigma \otimes I_n), \tag{1}$$

where: $X = (X_1, \ldots, X_n)_{sk \times n}$ is the response matrix, X_t $(t = 1, \ldots, n)$ being the response vector for all s sites at time t; $Z = (Z_1, \ldots, Z_n)_{h \times n}$ is the matrix of covariates;

$$B = \begin{pmatrix} \beta_{1,1} & \cdots & \beta_{1,h} \\ & \vdots & \\ \beta_{sk,1} & \cdots & \beta_{sk,h} \end{pmatrix}_{sk \times h}$$

is the coefficient matrix; Σ is the unknown spatial covariance matrix of X_t and I_n is a $n \times n$ identity matrix. The conjugate priors of Σ, B are,

$$B \mid B^*, \Sigma, F \sim N_{sk,h}(B^*, \Sigma \otimes F^{-1}) \tag{2}$$

and

$$\Sigma \mid \Phi, \delta^* \sim W_{sk}^{-1}(\Phi, \delta^*). \tag{3}$$

Among s sites, s_g are gauged and yield observations of pollutant concentrations. The remaining s_u ungauged sites provide no observations. Accordingly we partition X into X^0 and $X^{(1)}$. X^0 is the response matrix at ungauged sites. After appropriate rearrangement of columns, we can further partition $X^{(1)}$ into X^1 and X^2. The response matrix X^1 represents all unobserved pollutant concentrations and X^2 those observed. The partitions of Σ, B^*, F and Φ are similar. For instance,

$$\Sigma = \begin{pmatrix} \Sigma_{00} & \Sigma_{0(1)} \\ \Sigma_{(1)0} & \Sigma_{(11)} \end{pmatrix},$$

where Σ_{00} and $\Sigma_{(11)}$ are $s_u k \times s_u k$, $s_g k \times s_g k$ matrices, respectively.

An indicator matrix R is introduced to simplify notation. Suppose the indices of missing values in $X_t^{(1)}$ were $i_1, \ldots, i_l$ and the indices of observed values, $i_{l+1}, \ldots, i_{s_2 k}$. Let $R_1 = (r_{i_1}, \ldots, r_{i_l})$ and $R_2 = (r_{i_{l+1}}, \ldots, r_{i_{s_2 k}})$ where r_j, $j = 1, \ldots, s_2 k$ is a $s_2 k \times 1$-dimensional vector with the j^{th} element being one and the remainder being zero. Thus, R_1 and R_2 "mark" the position of missing columns. Then $R = (R_1, R_2)$, and we let $\Sigma_{ij} = R_i' \Sigma_{(11)} R_j$, $\Psi_{ij} = R_i' \Phi_{(11)} R_j$ and $B_i^* = R_i' B_{(1)}^*$, $i, j = 1, 2$.

For given hyperparameters, LSZ prove that the predictive distribution of $X^0 \mid X^2 = x^2$ follows a matrix T distribution. More precisely,

$$X^0 \mid X^2 = x^2 \sim T\left(\Phi_{0|2}^{-1}, c, B_0^* Z + \Phi_{0(1)} R_2 \Psi_{22}^{-1}(x^2 - B_2^* Z), \delta^* - I + 1\right)$$

where:

$$c = I + Z' F^{-1} Z + (x^2 - B_2^* Z)' \Psi_{22}^{-1}(x^2 - B_2^* Z);$$

$$\Phi_{0|2} = \Phi_{00} - \Phi_{0(1)} R_2 \Psi_{22}^{-1} R_2' \Phi_{(1)0}.$$

In the last result, I is the number of missing pollutant concentrations at gauged sites and time t. If we adopt a squared loss function, the Bayesian interpolator is

$$E(X^0 \mid X^2 = x^2) = B_0^* Z + \Phi_{0(1)} R_2 \Psi_{22}^{-1}(x^2 - B_2^* Z). \tag{4}$$

Following Brown, Le and Zidek (1994, hereafter BLZ), LSZ adopt an empirical approach and estimate hyperparameters. More precisely, LSZ maximize the conditional likelihood function for given $X^2 = x^2$ and also use two unbiased estimators. To reduce the number of parameters, LSZ adopt a Kronecker structure, $\Phi = \Lambda \otimes \Omega$, where Λ is the between-sites-hyperparameters and Ω, between-pollutants. LSZ use the following procedure to estimate all hyperparameters. First, they use two unbiased estimators to estimate $B_{(1)}^*$, F^{-1}; second, they apply the EM algorithm to estimate Ω, δ^* and Λ_g, where Λ_g has the between-gauged-sites-hypercovariance-matrix; third, they invoke a procedure of Sampson and Guttorp (1992, hereafter SG) to extend Λ_g to Λ. Finally, LSZ assume an exchangeability structure on B^* to extend $B_{(1)}^*$ to B^*.

SG's nonparametric approach estimates a spatial dispersion matrix when the data field is found to be anisotropic. The dispersion matrix has the same meaning as a variogram matrix except that isotropy is not implicitly implied. The SG method involves two steps.

First, with the nonmetric multidimensional scaling (MDS) algorithm (see Mardia, Kent and Bibby 1979), a two-dimensional representation of the sampling sites is found. In this two dimensional Euclidean space, called the D-plane, a monotonic function of the distance between two points approximates the spatial dispersion between the same two points. The D-plane, has a counterpart in the G-plane comprised of the geographical coordinates of the sampling sites. Step two yields thinplate splines to provide smooth mappings from the G-plane into their MDS representation. Then the composition of this mapping f and a monotone function g derived from MDS yields a nonparametric estimator of $var(Z(x_a) - Z(x_b, t))$ having the form $g(\mid f(x_a) - f(x_b) \mid)$ for any two geographic locations x_a and x_b. This rough g is then replaced.

Now let us turn to implementation issues in the next section.

3 Fitting the Interpolator

The daily maximum hourly levels of nitrogen dioxide (NO_2), ozone (O_3), sulphur dioxide (SO_2) and the daily mean levels of sulfate (SO_4) were recorded from January 1 of 1983 to December 31 of 1988 in Ontario and its surrounding areas. These data come from several monitoring networks in the Province, including the Environment Air Quality Monitoring Network (OME), Air Pollution in Ontario Study (APIOS) and the Canadian Acid and Precipitation Monitoring Network (CAPMON). The reader should see Burnett R. T. et al (1992) for a more detailed description of the data. In all, the network has 37 different monitoring locations (sites) but not all sites monitor all of the four air pollutants.

Monthly average pollutant concentrations at gauged sites are simply computed as the mean of the observed daily levels for that month, January 1983 to December 1988. The time series of observed monthly mean concentrations for each pollutant consists of 72 values. The series with more than one third missing values are omitted from this analysis. As a result, the number of gauged sites is reduced to 31 from 37. Figures 3 depicts the locations of each pollutant measured at a subset of the remained 31 sites. The whole Ontario Province divides into thirty-seven PHUs or districts (Duddek et al 1994). The PHU is similar to a Census Division, the difference being marginal disagreements in boundaries. Some PHUs, for example, are aggregates of two Census Divisions. Figure 4 displays the locations of the approximate centroids of these PHU's. Hence, the total number of gauged sites s_g is 31 and

the total number of ungauged sites s_u is 37.

At the 31 gauged sites, there are 64 observed and 60 missing time series. Among the 64 observed time series, about two percent of the values are missing, including those below the detection limit. In this analysis, each of the missing values is replaced by the mean of the monthly observed values of the same pollutant and month in other years. If all six measurements in the same month are missing, the grand mean of observations in the six years will be used. However, no such case exists in the data set. A more delicate method of filling in the missing data may be used here. But with such a low percentage of missing data the extra effort seems unnecessary.

The LSZ theory is developed under two important assumptions of normality and temporal independence (see Equation (1)). Checking multivariate normality assumption is not easy. In this paper, we only examine the normal quantile plot for each pollutant separately. The normal quantile plots of the residuals of the raw data seem very nonlinear. Therefore, the observed data must be transformed. With a logarithmic transformation of the observed data, the residuals appear to be marginally normal. Figure 5 shows a typical example of the normal quantile plots of the data. The plot is based on the measured air pollution levels of SO_4 at gauged Site 4. In the sequel, when we refer to these pollutants we mean their log-transformed versions.

The temporal independence assumption is checked with autocorrelation and partial autocorrelation plots. Autocorrelation and partial autocorrelation plots of the temporal residuals of the log-transformed data are shown in Figure 7. The plot is based on the measured air pollution levels of SO_4 at gauged Site 1. The correlation plots show no sign of autocorrelation. By repeating the above initial data analysis for the observed pollutant levels at all gauged sites, we conclude that the assumptions of our interpolation theory seem reasonable with the log-transformed data.

The linear trend and seasonal component of the time series are captured with $Z_t = \{1, t, \cos(2\ \pi\ t/12),\ sin(2\pi t/12)\}$, where $t = 1, \ldots 72$. Here $t = 1$ represents the January of 1983, $t = 2$ represents the February of 1983 and so on, until $t = 72$, which represents December of 1988. The coefficients of the linear trend and seasonal component are estimated with ordinary least squares. In Figure 6, the time series plots and the least squares fitted curve of the four observed pollutants at Site 5 are displayed. The fit of the time series for

	NO_2	SO_4	O_3	SO_2
NO_2	1.00	-0.29	0.03	0.14
SO_4	-0.29	1.00	0.79	-0.34
O_3	0.03	0.79	1.00	-0.15
SO_2	0.14	-0.34	-0.15	1.00

Table 1: The Estimated Between-pollutants-hypercorrelation Matrix of the Log-transformed Monthly Data

$log(O_3)$ is far better than that of the other three, because of its periodic nature. The strong yearly pattern of ozone is partially explained by the fact that the creation of ozone is highly related to solar radiation.

The air pollution level in summer is of special interest in health impact analysis (not shown). In the following, only the interpolation for summer data, i.e. from May 1 to August 31, is demonstrated. Each summer data time series thus has 24 values (months). We take as our purpose, the interpolation down to 37 PHU approximate centroids in Southern Ontario, of NO_2, SO_4, O_3 and SO_2 levels in the summers of 1983 to 1988.

The interpolation procedure begins by finding the unbiased estimators of F^{-1} and $B^*_{(1)}$; next, the EM algorithm is invoked to estimate δ^*, Λ_g and Ω; third, the SG method is applied to extend Λ_g to Λ; then, with the exchangeability assumption on B^*, $B^*_{(1)}$ is extended to B^*; finally, all hyperparameters having being estimated, the interpolated values are computed by the Bayesian interpolator.

Software to implement the approach has been developed and a working version is now available. Applying the approach to the summer data yields an estimate of 610 for the prior number of degrees of freedom. Table 1 gives the corresponding estimate of the hyper-correlation matrix of the log transformed NO_2, SO_4, O_3 and SO_2 values; the corresponding hyper-variances are 0.66, 1.63, 0.22, 1.85. Among the estimated hyper-variances, that of $log(O_3)$ is smallest, $log(SO_2)$, largest. This result indicates that the overall variation of the observed ozone levels is smaller than that of SO_2. The result confirms our prior knowledge that ozone unlike SO_2 is a regional pollutant and so more homogeneous. The biggest pairwise correlation among the four pollutants is found to be between O_3 and SO_4. Since both O_3 and SO_4 are regional air pollutants and both are related to sunlight, we would anticipate that result.

Figures 8, 9 and 10 summarize the result of the SG step. The righthand plot in Figure 8 is a twisted 30-by-30 checkerboard in the D-plane. The original 30-by-30 checkerboard is in the G-plane and the coordinates of its lower left corner are the minimum latitude and longitude of the gauged sites. The coordinates of its upper right corner are the corresponding maximum. The lefthand plot in Figure 8 shows an exponential fit between dispersion and the D-plane distance (refer to Section 1 for a brief summary of the SG method). The parameter λ controls the smoothness of the twisted checkerboard. By sacrificing the fit between the dispersion and D-plane distance, we get a flatter checkerboard. Figure 9 shows that checkerboard along with a rougher fit between the dispersion and D-plane distance obtained when the smoothing parameter value increases from 0 to 2500. The straight line in the righthand side of Figure 10 shows that the estimated covariance and the observed covariance are conformable.

After applying the GS method, we compute the interpolated air pollutant levels at all the PHU approximate centroids over six years by applying Equation (4) and using the above estimated hyperparameter values. To check the interpolated values, we plot in Figure 11 the overall means of observed ozone levels at each gauged site in the summers of 1983 to 1986. Those of interpolated ozone levels at the PHU approximate centroids appear in Figure 12. As expected, when a higher mean O_3 level is observed at a gauged site, our Bayesian method interpolates higher O_3 values at the PHU approximate centroids near that site. Analogous results obtain for a lower observed O_3 level.

4 How Accurate is the Interpolator?

In this section we study the performance of the interpolator fitted in the last section.

We can assess the interpolation procedure by looking at the correlation between the observed and estimated data through cross validation (CV hereafter). CV successively deletes monitoring sites one at a time and imputes their missing data from the remainder. In our study, we deleted one gauged site at a time and interpolated the pollutant levels at the same site using the observed levels at the other sites.

Table 2 gives the correlation between the estimated and observed levels at each gauged site for each observed pollutant. Notice that the correlations between SO_4 and O_3 are generally higher than those of SO_2. This finding suggests that predicting SO_4 or O_3 is easier

Sites	NO2	SO4	O3	SO2
1		0.99		
2		0.98		
3		0.96		
4		0.97		
5	0.58	0.86	0.97	0.61
6			0.98	
7		0.98		0.88
8	0.76		0.95	0.81
9			0.99	
10			0.90	
11	0.64		0.99	0.72
12			0.98	0.91
13	0.56		0.94	0.64
14	0.39		0.96	0.61
15	0.77		0.98	0.72
16		0.99		
17	0.73	0.97	0.97	0.91
18		0.98		
19	0.68		0.95	0.81
20	0.39		0.93	0.88
21	0.69		0.95	0.68
22			0.98	
23				0.91
24	0.66		0.98	0.66
25			0.96	0.57
26	0.75		0.93	0.90
27	0.76		0.94	0.63
28			0.95	0.91
29		0.97		
30				0.90
31			0.91	0.86
Mean	0.64	0.97	0.96	0.78

Table 2: Correlations Between the Log-transformed Observed and Estimated Pollution Levels at Gauged Sites

than SO_2.

Figure 13 confirms that finding. That figure displays the plot of residuals of log-transformed monthly observed and estimated pollutant levels.

Similar findings are revealed in Figure 14 which plots estimated levels against observed levels for each (log-transformed) pollutant. In those plots, a straight line would mean perfectly accurate interpolation. The plots confirm conclusions suggested by the tables and demonstrate again that O_3 and SO_4 are regional pollutants. They are easier to predict than their nonregional counterparts.

In the analysis described above, we explored the quality of the interpolator (posterior mean). However, one potentially important advantage of the LSZ method lies in its capability to provide a predictive distribution. We need to assess the quality of the predictive distribution as a whole to fully determine the value of the method.

Note that the marginal distribution of a Matrix T is another matrix T (c.f. Press 1982); when the numbers of its rows and columns are one, the matrix T becomes a student's t. By repeatly applying this fact, we can theoretically derive the univariate predictive distribution of any given pollutant at an ungauged site at any specified fixed time. Finding a 95% credibility interval for that univariate (t) distribution is easy. The quality of that distribution can then be judged by seeing how often the observations fall inside these intervals. We apply this idea to the CV study described above.

First, at a "deleted" gauged site, we compute a 95% credibility interval for each of the four observed pollutants and of the 24 summer months. Second, we count how many of these intervals cover the observed values and find the coverage percentages for NO_2, SO_4, O_3 and SO_2, respectively. These percentages are respectively, 88.1, 97.5, 98.8 and 99.6. These percentages deviate from 100%, some appreciably. These deviations may be explained by the large number of degrees of freedom (m) selected by the EM algorithm for the joint distribution of the four pollutants. For NO_2 especially, the resulting marginal distribution has excessively light (normal-like) tails. We learn from this analysis of the need, in marginal analysis to select m differently. Indeed, for NO_2 we find through CV assessment, that m would have to be about 10 to give the required 100% coverage. However, our results remain too incomplete to be presented in this paper.

The analysis of the last paragraph bears on the following question. Can a simpler to use, normal distribution be substituted for the multivariate T predictive distribution? That might naively seem possible since the univariate normal approximates its longer tailed relative quite well. However, our results suggest this substitution cannot be recommended at least without additional study. Our initial impression comes from an evaluation we did of the empirical coverage percentage of three-standard-deviation confidence intervals (CI). If the predictive distribution were normal, all the three-standard-deviation CIs would include the true values about 100 percent of the time. As the percentages in Table 3 indicate, this high coverage probability is not achieved, particularly for SO_2. The heavier tailed predictive matrix T distribution seems to be required.

Pollutant	Coverage
NO_2	100%
SO_4	100%
O_3	98.6%
SO_2	94.5%

Table 3: Empirical Percentages of Three SD's Intervals

Besides checking the empirical coverage probabilities for the pollutant-wise marginal credibility intervals, we also looked at them for the simultaneous case. LSZ give the formula for a simultaneous credibility region at level $1 - \alpha$. With it, at a fixed summer month and "deleted" gauged site, we test whether the observed vector fall inside the corresponding credibility region (hyperellipsoid). Because parts of individual site response vector are missing-by-design, the lengths of observed vectors at different "deleted" gauged sites may differ and so might their hyperellipsoids. The empirical coverage probability equals to the proportion of the observed vectors in their hyperellipsoids. At levels of 50%, 80%, 90%, 95% and 99%, they are 57.2%, 81.7%, 89.5%, 94% and 98%, respectively.

5 Multivariate vs. Univariate Interpolation

By interpolating one pollutant at a time, one can apply the univariate theory, proposed by Le and Zidek (1992), to the problem studied above. So why a new theory when an old theory exists? The answer lies in the information gained in the new approach and the corresponding increase in the accuracy.

With the univariate method, only partial data are used for each interpolation. The new method includes all the available data. In this section, we compare the performance of the two methods: *univariate interpolation* and *multivariate interpolation*. Consider the interpolation of O_3 in the southern Ontario study for example. When the levels of O_3 are interpolated down to the ungauged sites by the univariate theory, only the observed O_3 levels at gauged sites are included in the interpolator. By the new method, all observed values of NO_2, SO_4, O_3 and SO_2 are included.

Theoretically we can show that the multivariate interpolator leads to a smaller mean square error than that of its counterpart. More precisely, let X_0, Y_0 be any two random vectors and X a random variable. Then

$$E(X - E(X \mid \mathrm{X}_0, \mathrm{Y}_0))^2 \leq E(X - E(X \mid \mathrm{X}_0))^2. \tag{5}$$

The proof can be found in Sun (1994).

Returning to the O_3 example, we take X_0 to be the observed levels of O_3 at the gauged sites, Y_0, the observed levels of the other pollutants and X, any unobserved pollution level at an ungauged site. Then the univariate Bayesian interpolator is, $E(X \mid \mathrm{X}_0 = x_0)$, the multivariate, $E(X \mid \mathrm{X}_0 = x_0, \mathrm{Y}_0 = y_0)$. When the model is correctly specified and all the hyperparameters are known, the theoretical result above implies that the multivariate interpolator does at least as well as the univariate one.

The following CV study answers empirically the same question, again, using the monthly air pollution data set from southern Ontario. At each gauged site successively, the observed pollutants are deleted as if they were not observed. Then both univariate and multivariate Bayesian interpolators are applied to obtain the predicted values of the "deleted" values based on the data at the other gauged sites. When the values are predicted by both methods for all 31 gauged sites, we calculate the mean squared prediction error for the univariate and the multivariate interpolator, respectively. The results for the monthly summer data are listed below.

The values depicted in Table 4 confirm the theory.

One interesting point bears emphasis. Our results show that the relative reduction of the mean squared prediction error from using multivariate interpolation over univariate inter-

Pollutant	Multivariate	Univariate
NO_2	0.19	0.28
SO_4	0.14	1.27
O_3	0.05	0.13
SO_2	0.62	0.76

Table 4: Mean Squared Prediction Error for Multivariate and Univariate Interpolator

polation is much higher for SO_4 and O_3 than SO_2. For SO_4 and O_3, the changes are from 1.27 and 0.13 to 0.14 and 0.05 respectively; for SO_2 and NO_2, from 0.76 and 0.28 to 0.62 and 0.19 respectively. This result like others in Section 3, can be explained by the fact that SO_4, O_3 are regional pollutants while SO_2 and NO_2 are not. A regional air pollutant has higher correlation with other pollutants, as indicated by the estimated between-pollutants-hypercorrelation in the previous Section. Including the other correlated pollutants in the analysis should enhance the interpolator's performance relatively more. For a local pollutant, since it has little or no correlation with other pollutants, the inclusion of additional pollutants in the analysis will not improve the interpolator as much. Therefore, we can conclude on heuristics alone that the multivariate interpolator does better than the univariate interpolator. It does not do so much better on local pollutants however.

To further strengthen our comparison, we use *log predictive scoring*. First, we compute the mean log score for univariate interpolation. In particular, we: choose a pollutant; delete a gauged site for that pollutant; compute the predictive density function for the pollutant at a fixed summer month; plug the "deleted" value into the density function to get the log predictive score and finish the above process for each summer month; change to the next pollutant and repeat the same thing. The log predictive score for univariate interpolation is the grand mean of those log scores. That value is -8.41. Second, for the log predictive score from multivariate interpolation, deleting a gauged site at a time, we: compute the marginal predictive density functions for each pollutant at a fixed summer month based on the simultaneous predictive density function; find the log predictive density (score) for each observed monthly pollutant level and take the mean of the log scores. We found the value to be -0.68. Since the log predictive score explains how well a particular "observed" value is predicted by its predictive distribution, a bigger score by a predictive function implies stronger predictability. Based on those mean log scores, we conclude that the univariate marginal predictor derived from the multivariate one does better than the completely univariate one.

References

[1] Brown, P.J., Le, N. D. and Zidek, J.V. (1994). "Multivariate Spatial Interpolation and Exposure to Air Pollutants". Canadian Journal of Statistics. **22**, 489-509.

[2] Burnett, R. T., Dales R. E., Rainenne M. D. and Krewski D. (1992). "The Relationship Between Hospital Admissions and Ambient Air Pollution in Ontario, Canada: A Preliminary Report". Unpublished Report.

[3] Le, N. D., Sun, W. and Zidek, J.V. (1994). "Bayesian Multivariate Spatial Interpolation with Data Missing-by-design." Submitted.

[4] Le, N. D. and Zidek, J.V. (1992). "Interpolation with Uncertain Spatial Covariance: A Bayesian Alternative to Kriging". J. Mult. Anal, 43, 351-74.

[5] Press, S. J., (1982). "Applied Multivariate Analysis-Using Bayesian and Frequentist Methods of Inference". Holt, Rinehart & Winston, New York.

[6] Sampson, P. and Guttorp, P., (1992). "Nonparametric estimation of nonstationary spatial covariance structure". *J. Amer. Statist. Assoc.* Vol.87 No. 417, 108-119.

[7] Sun, W. (1994). "Bayesian Multivariate Interpolation with Missing Data and Its Applications". Unpublished Ph.D. Thesis, Department of Statistics, University of British Columbia, Vancouver, Canada.

Appendix: Figures

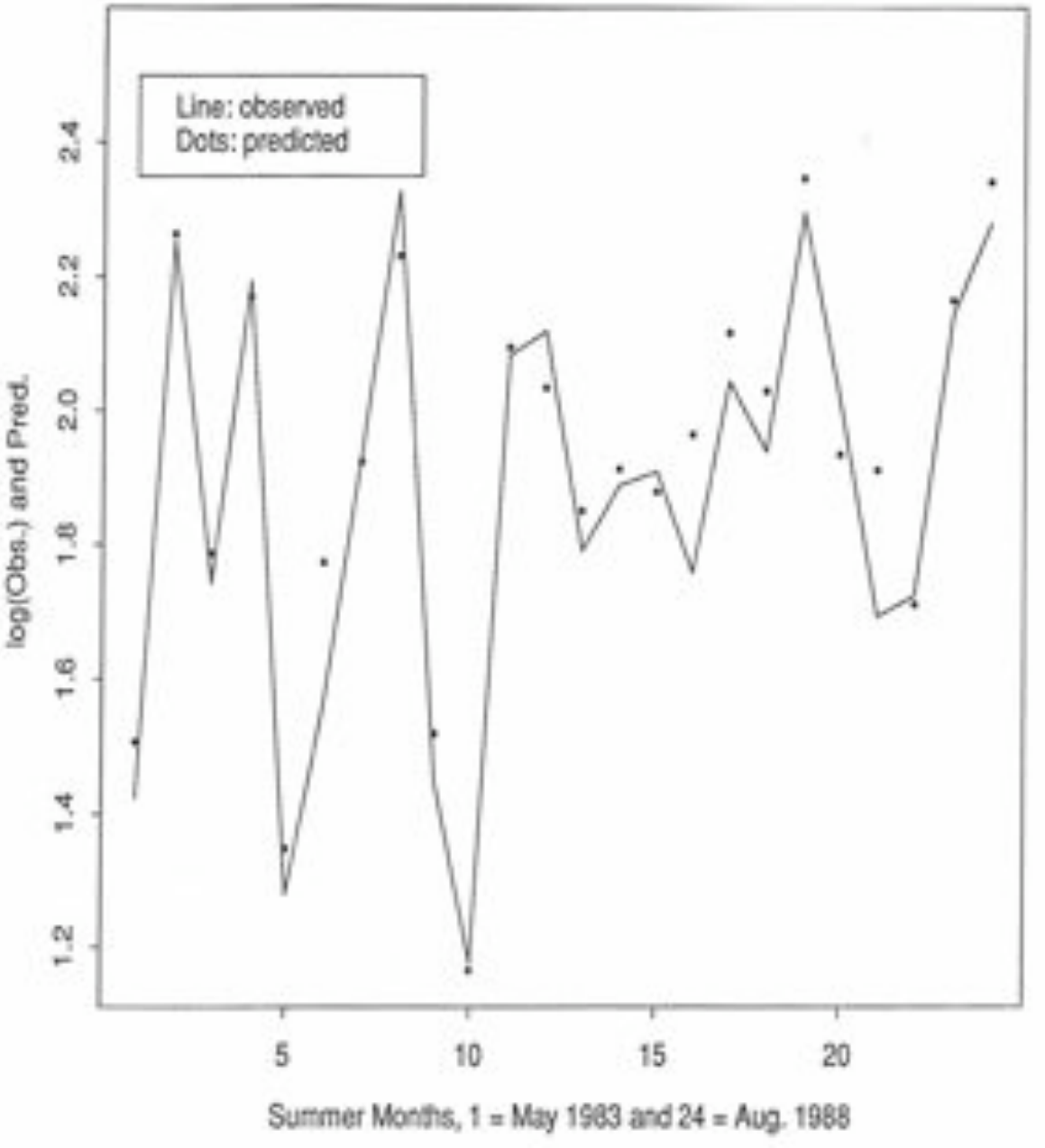

Figure 1: Observed and predicted values of SO_4 at Site 17 (26 Breadalane Toronto).

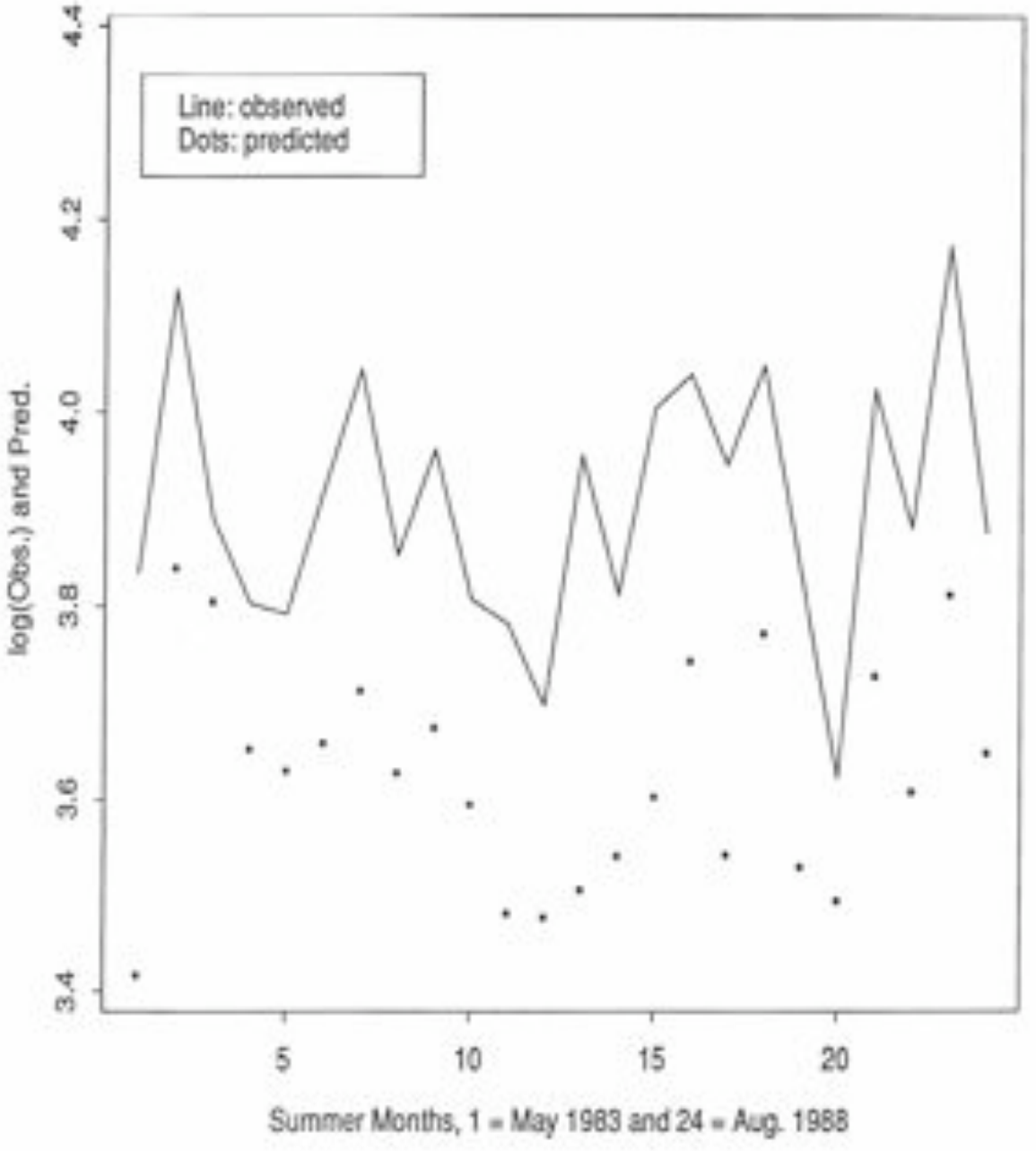

Figure 2: Observed and predicted values of NO_2 at site 17 (26 Breadalane Toronto)

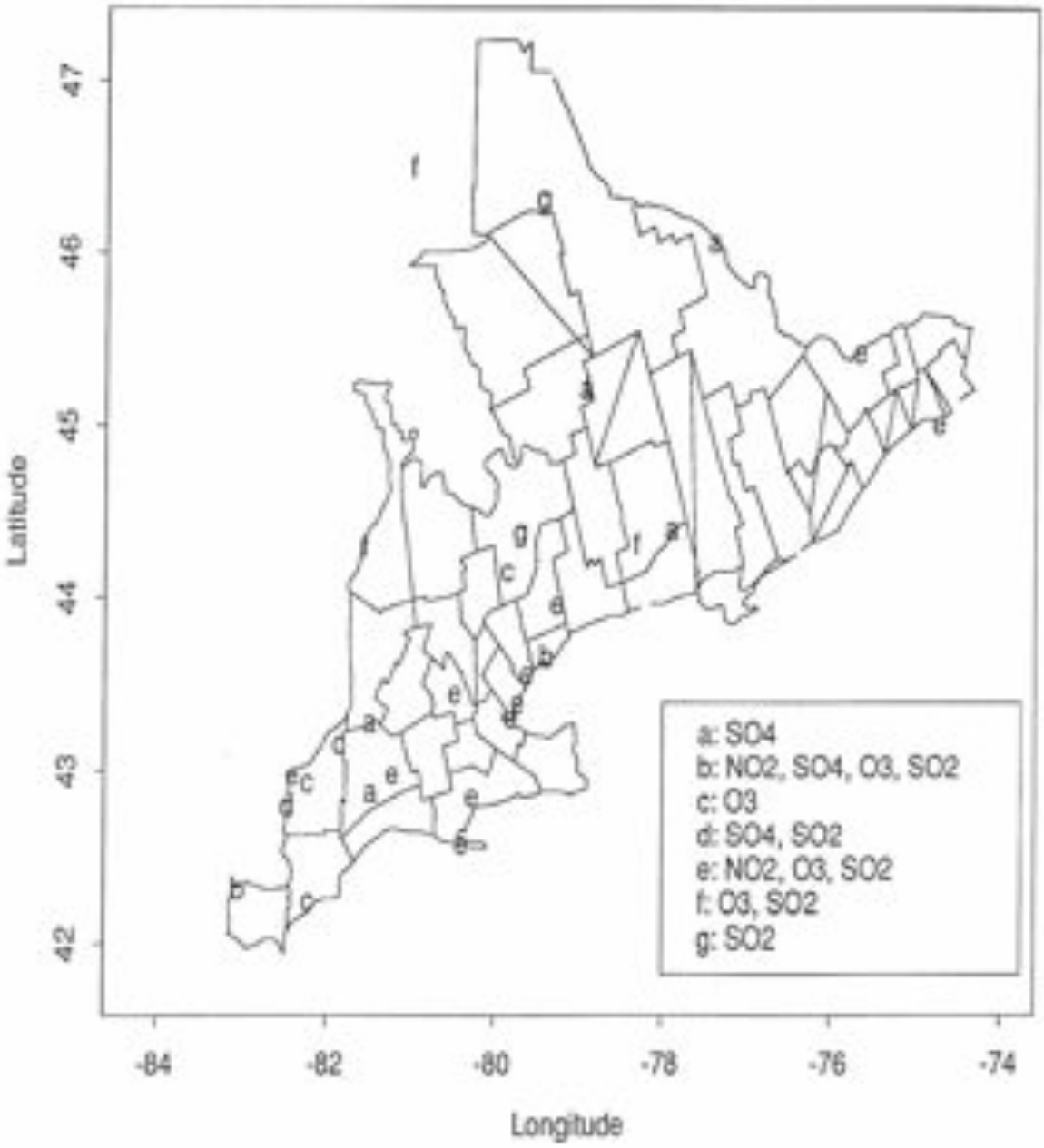

Figure 3: Monitoring Sites Yielding SO_4 Data

Figure 4: Locations of selected sites in Southern Ontario plotted with Census Subdivision's boundaries, where monthly interpolated pollution levels are needed.

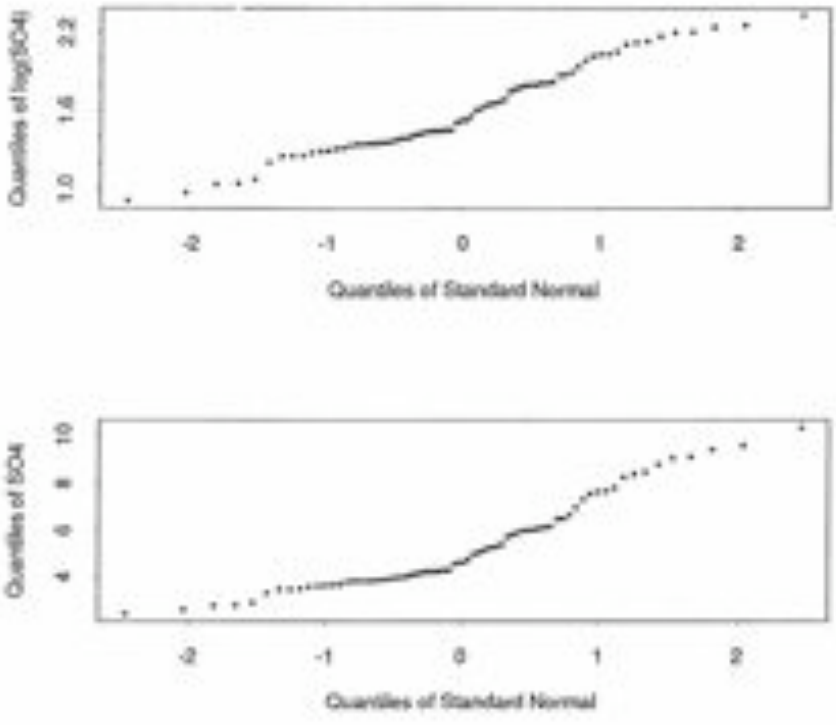

Figure 5: Normal quantile-quantile plots for original and log-transformed monthly levels of SO_4 in $\mu g/m^3$ at Gauged Site 4.

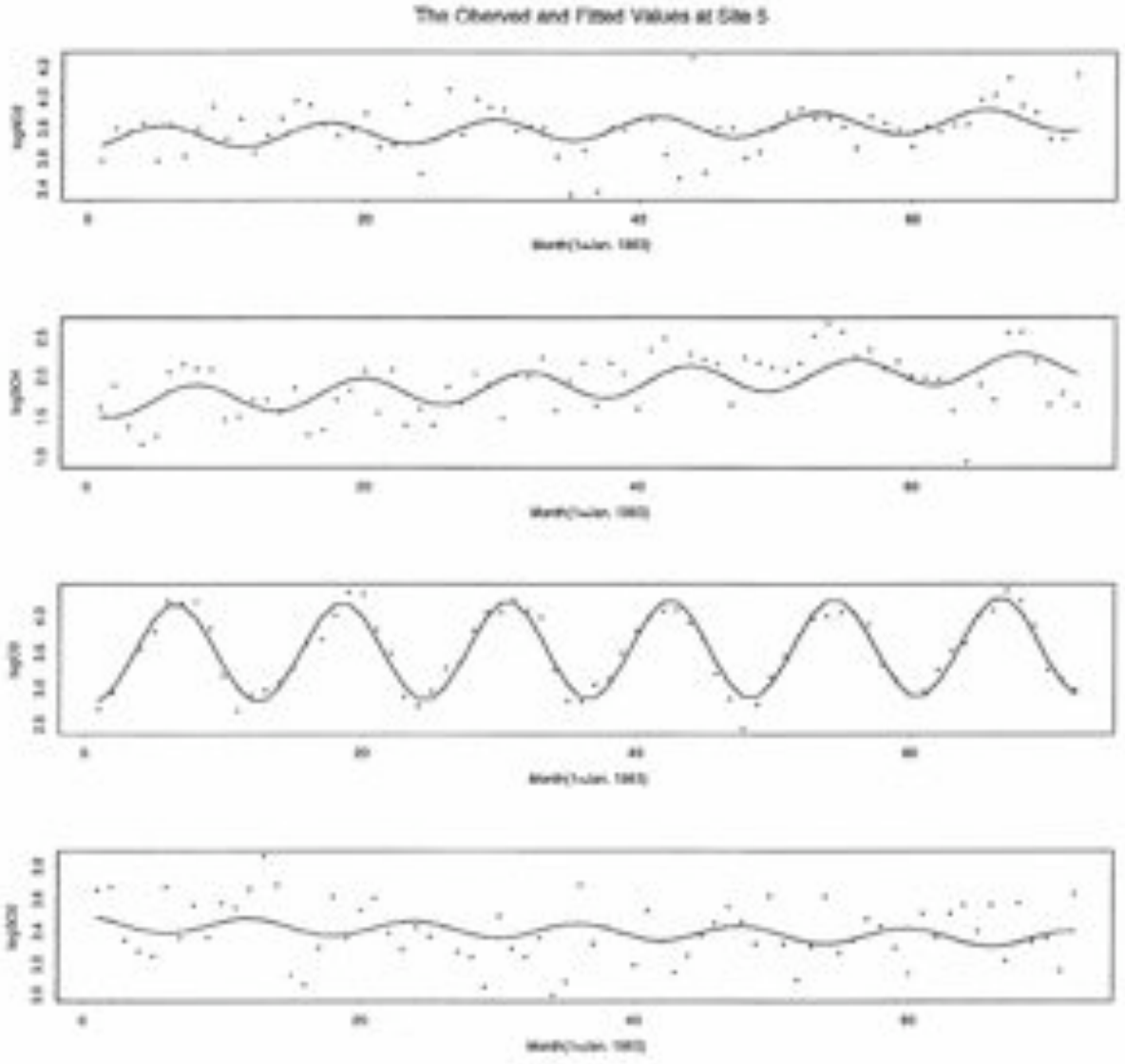

Figure 6: Plots for monthly observed and fitted, log-transformed levels of O_3 in ppb, SO_2, NO_2 and SO_4 in $\mu g/m^3$, at Gauged Site 5.

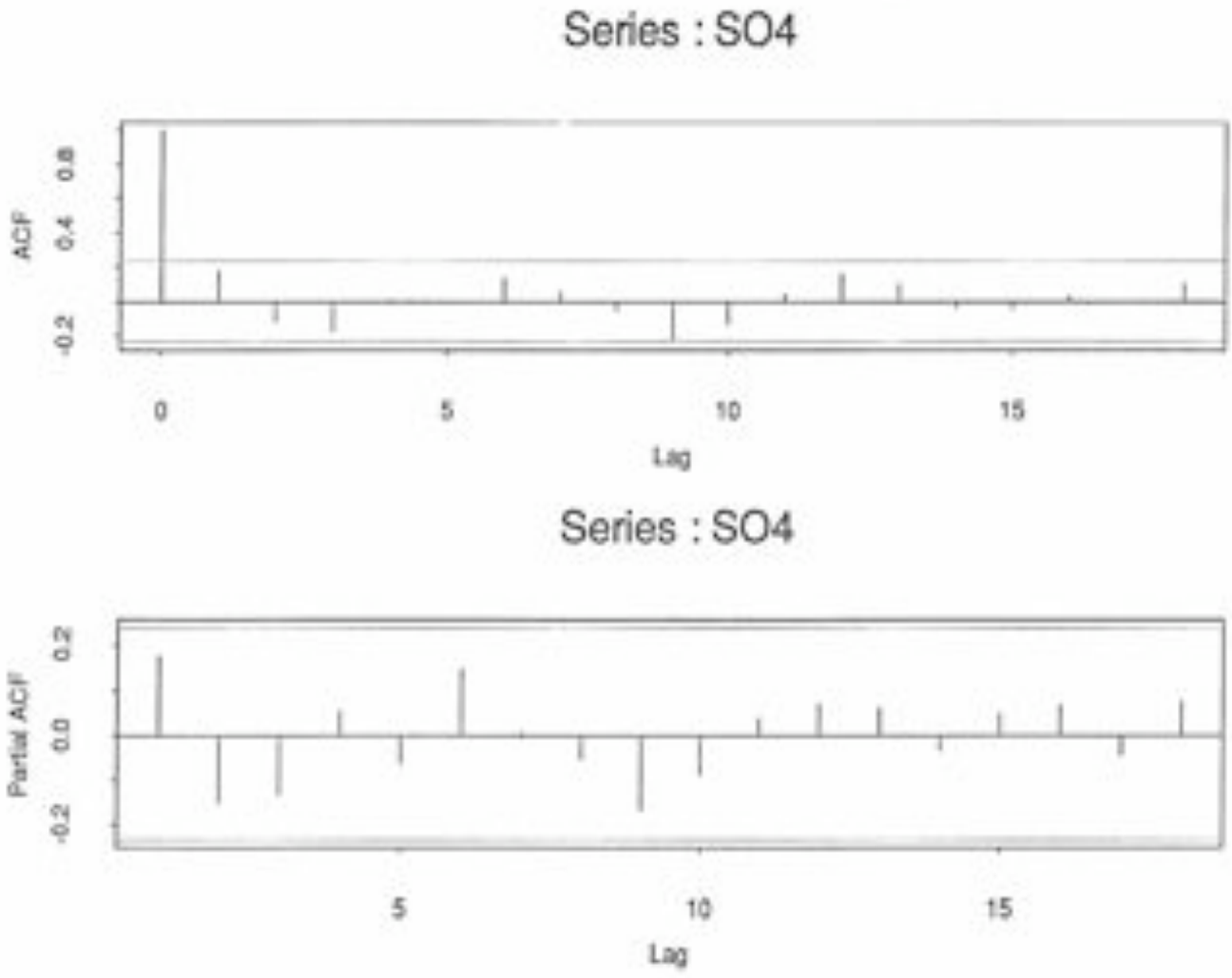

Figure 7: Plots for autocorrelation and partial autocorrelation of monthly, log-transformed levels of SO_4 in $\mu g/m^3$ at Gauged Site 4.

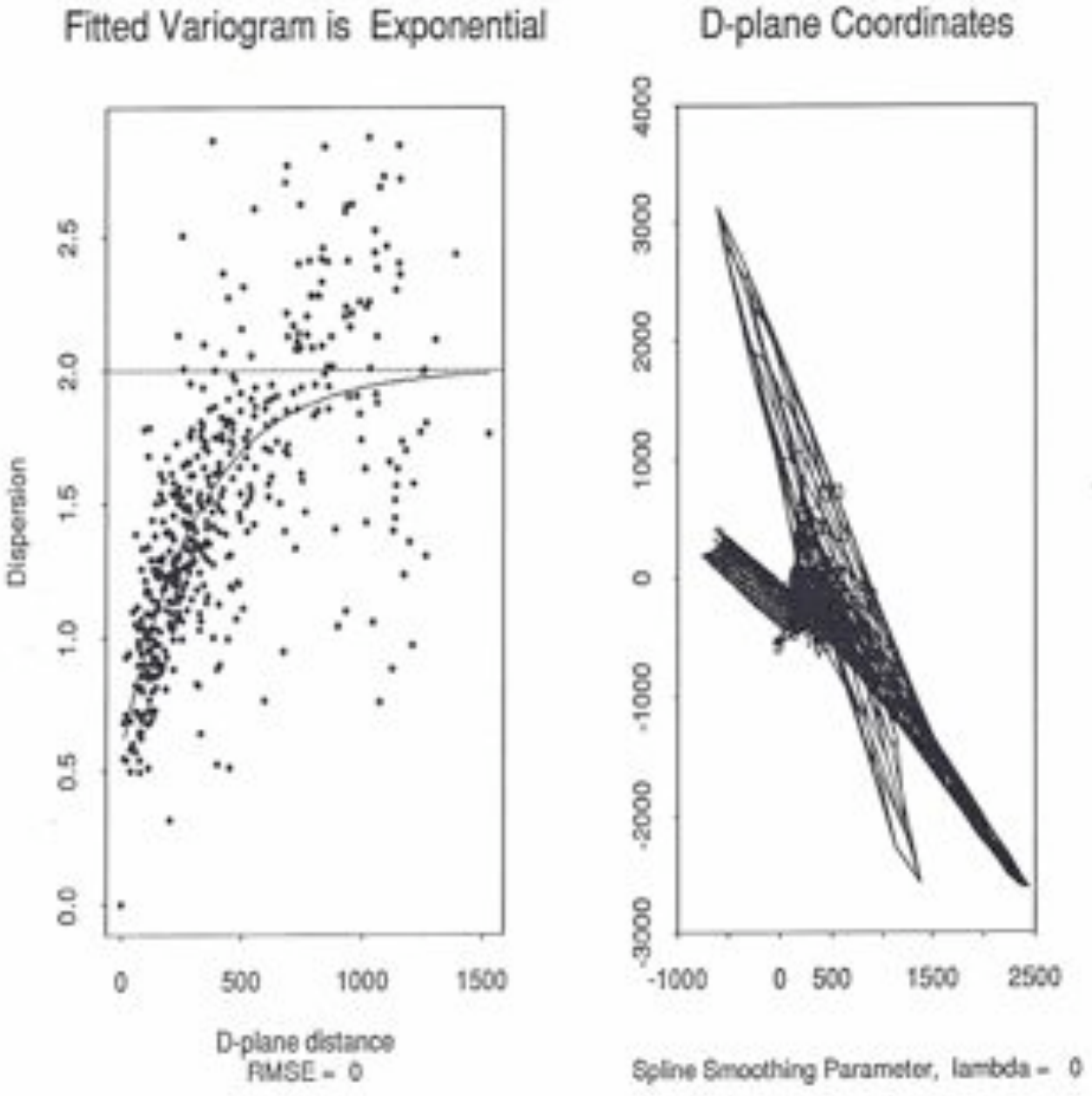

Figure 8: A rough checkerboard obtained in the SG step.

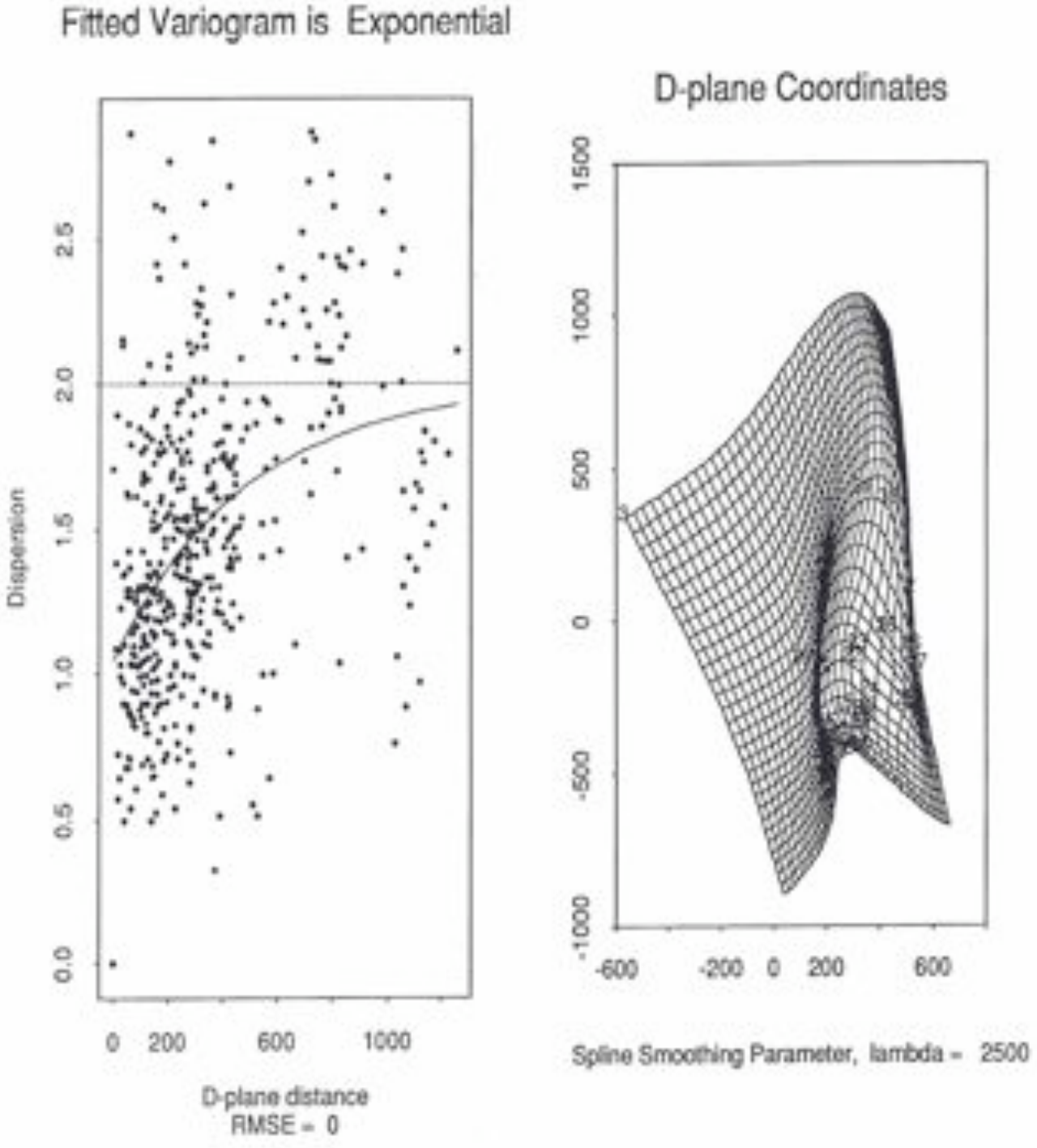

Figure 9: A smoother checkerboard obtained in the SG step.

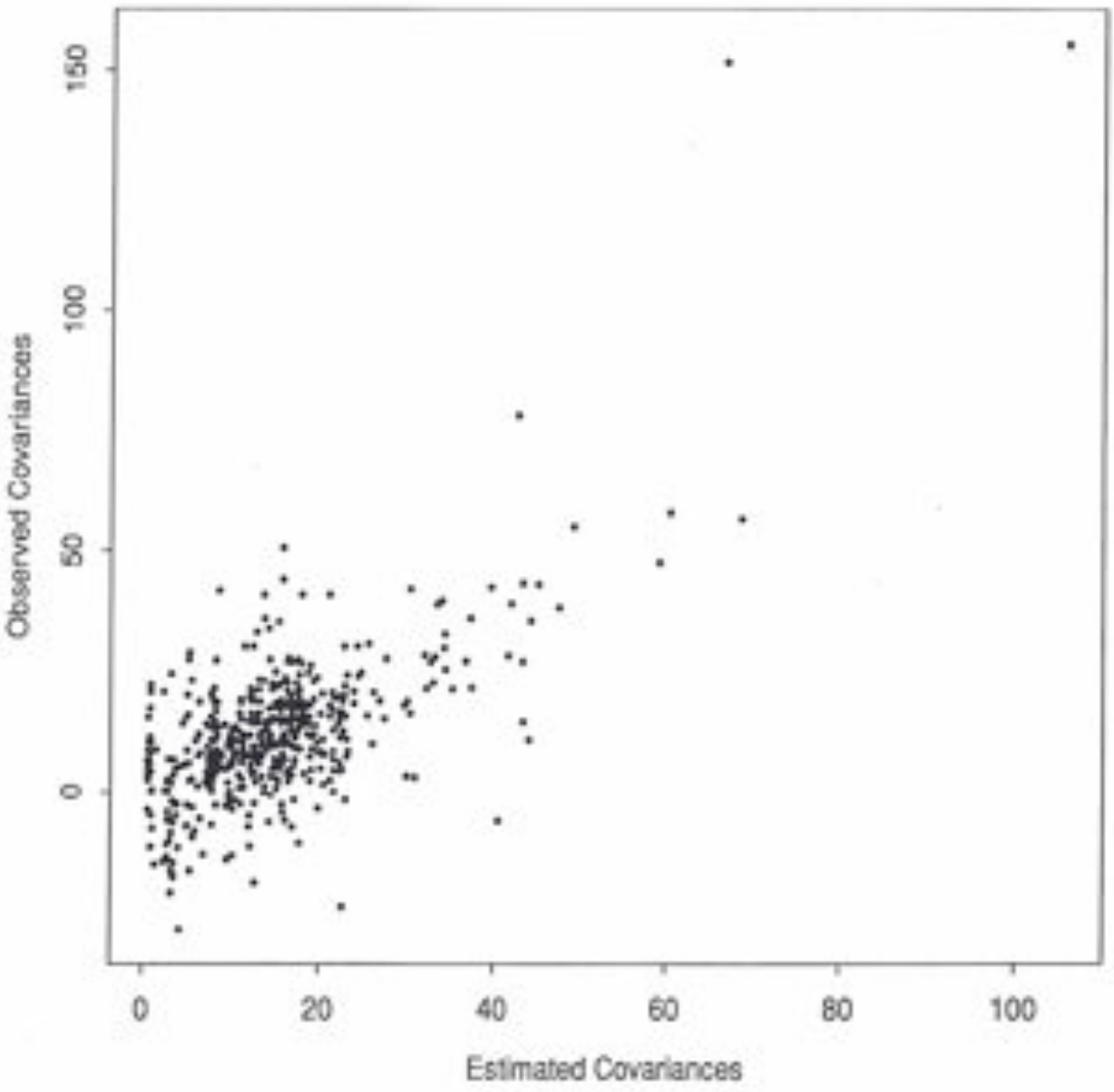

Figure 10: Scatter plot of observed covariances vs predicted covariances obtained by the GS approach.

Figure 11: Means of monthly levels of O_3 in ppb, in summers of 1983 $\sim$ 1988 at gauged sites in Southern Ontario plotted with CSD boundaries.

Figure 12: Means of monthly levels of O_3 in ppb, in summers of 1983 $\sim$ 1988 at selected sites in Southern Ontario plotted with CSD boundaries.

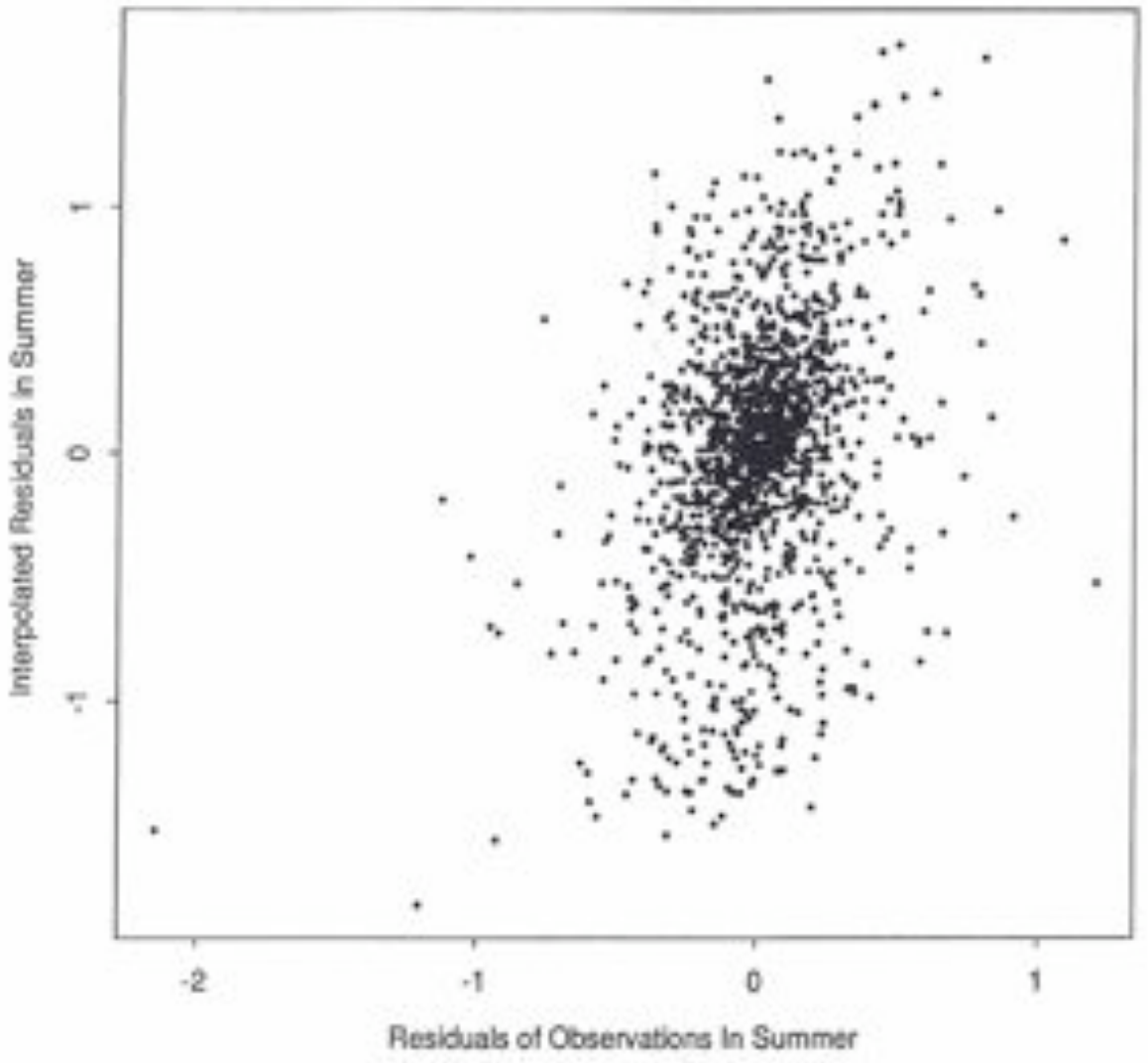

Figure 13: Scatter plots for residuals of monthly observed pollutant levels against residuals of interpolated levels at the log-scale in summer, where levels of O_3 are in *ppb*; SO_2, NO_2 and SO_4 in $\mu g/m^3$.

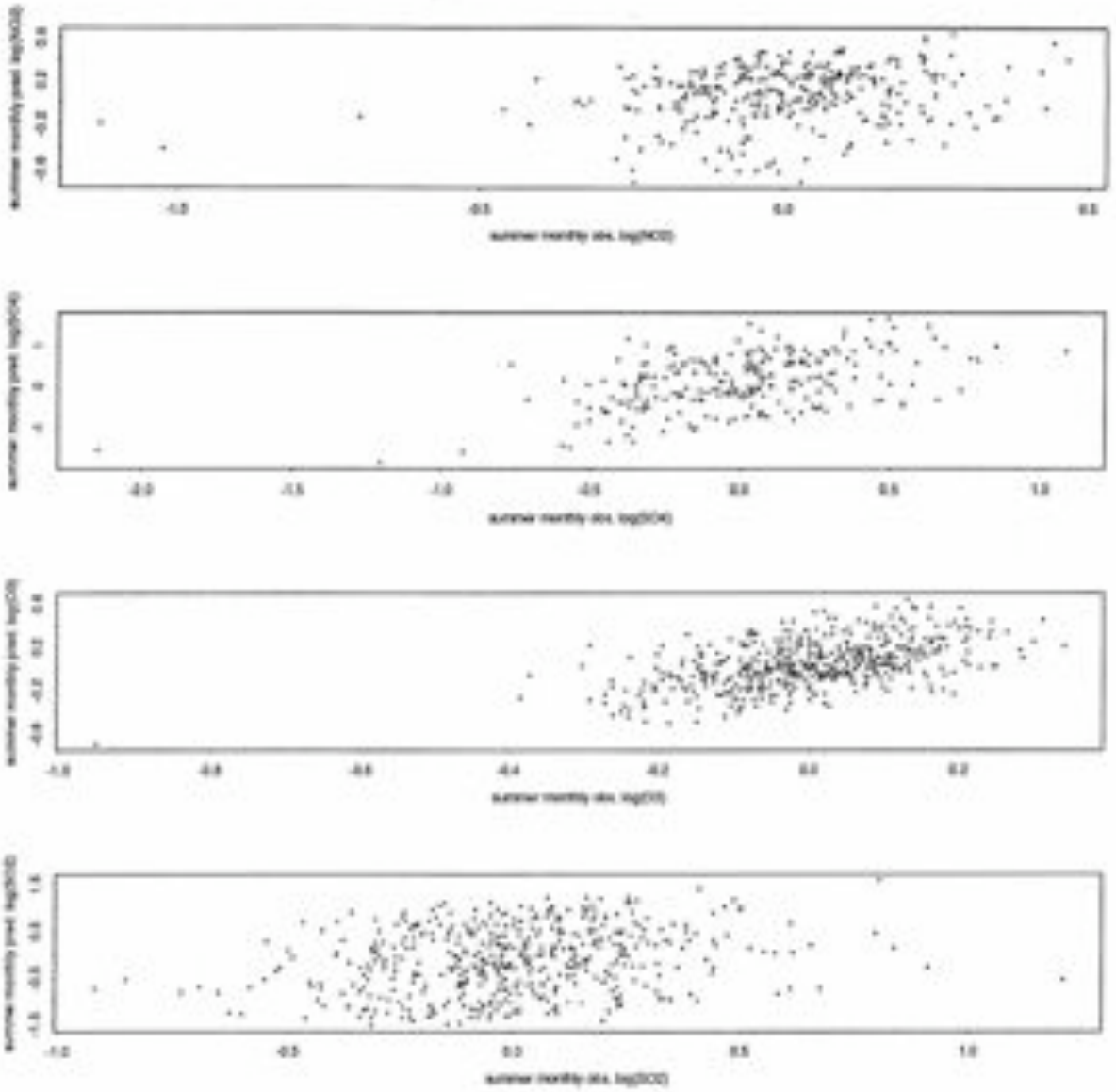

Figure 14: Pollutant-wise scatter plots for residuals of monthly observed pollutant levels against residuals of interpolated levels at the log-scale in summer, where levels of O_3 are in *ppb*, SO_2, NO_2 and SO_4 in $\mu g/m^3$.