Spatio-Temporal Modelling of the Longitudinal $PM_{2.5}$ Data with Missing Values

Stanislav Kolenikov* Richard L. Smith

University of North Carolina at Chapel Hill
Draft — not for citation... yet...

Abstract

Abstract: This paper analyzes the data on the particular matter of size of 2.5 microns or less ($PM_{2.5}$) for the states of North Carolina, South Carolina and Georgia. The spatio-temporal model features additive semi-parametric spatial and temporal trends, i.e., thin-plate splines, and B-splines, respectively. Reduction in the number of nodes for the thin-plate component is achieved through the use of clustering routines that combine adjacent observations sites. The residual spatial covariance is modelled by an exponential-power family semivariogram with nugget effect. The parameters of the model are estimated by the generalized EM algorithm which allows to account for the missing data. The relation between predictions produced by the EM algorithm and the universal kriging is discussed.

The resulting PM maps suggest the anthropogenic origin of the particular matter. In particular, the $PM_{2.5}$ concentrations are lower off the coast and in sparsely inhabited Appalachians, although the standard errors of predictions for those regions are higher, as those are rather remote from most monitors. The concentrations of the $PM_{2.5}$ were found to have stayed around 10-15 $\mu g/m^3$ for most of the year, and increase to 25-35 $\mu g/m^3$ in late summer. It was also found that most of Georgia violates the federal regulations of 15 $\mu g/m^3$ most of the time.

1 Introduction

Airborne particular matter has become an important topic of epidemiological and environmental studies in the last ten or so years when it was understood that

*Correspondence to 117 New West Bldg., Cameron Ave, CB # 3260, Chapel Hill, NC 27514, or skolenik@unc.edu.

the particular matter is a fairly important determinant of the deaths, especially in the elderly, even though the biological mechanisms of its effect are not quite clear yet. The United States Environmental Protection Agency regulates the admissible levels of PM_{10} and $PM_{2.5}$, the indicators of the concentration of the particular matter of sizes 10 and 2.5 μm, respectively[1]. The federal standard for $PM_{2.5}$, the particular matter size studied in this paper, was introduced in 1996, and states that "(a) the three-year average of the annual 98th percentile of $PM_{2.5}$ concentration measurements at any monitoring site should not exceed 50 μg per cubic meter of ambient air, and (b) the arithmetic mean (over one or multiple monitoring sites in the region) of site-specific three-year averages of daily $PM_{2.5}$ concentration measurements should not exceed 15 μg per cubic meter" (Cox 2000).

The EPA has also outlined a number of research topics related to the particular matter, and one of the statistical questions raised is, "Can spatial interpolation methods provide more accurate estimates of individual exposures to particulate air pollution?" (Cox 2000).

The development of spatial statistical models began in the 1950s in mining and agricultural applications. Most of the developments in spatial methods are summarized in Cressie (1993). Construction of the spatio-temporal models has become popular rather recently, in the 1990s.

Probably the most general framework for spatio-temporal modeling is outlined by Haas (2002). He proposes to first transform the dependent variable towards normality by extending its range over the whole real line and reducing its skewness and kurtosis to zero by suitable power-type transformations. He then defines two main basic approaches to spatio-temporal modeling as the local modeling and the global modeling. The former approach, called LOMAP in Haas (2002), estimates the model parameters and constructs predictions for a relatively small number of observations within the *prediction cylinder* defined as the observations close to the given one both in time and in space. The distance in the three dimensional space-time construct is a lexicographic one: the observations are first sorted by their distance in the spatial plane (spatial lag), and then by the distance in time (temporal lag). (An earlier version of this local approach is given in Haas (1998) under the name of *moving cylinder spatio-temporal kriging*.). There can be many such cylinders if the quantity of interest is multivariate. The model is estimated by the Cramér-von Mises distance between the empirical distribution function and the standard normal CDF (*minimum distance* method). Finally, the point prediction and its standard error are approximated by Monte Carlo simulation from the Gaussian distribution

[1]The precise definition of the $PM_{2.5}$ is the particle size at which 50 per cent of the particles of this size (aerodynamic diameter) are collected by the monitoring device Cox (2000).

with the estimated parameters[2]. In the global modeling approach (GLOMAP), the overall model is obtained as a mixture of the component models with smooth weights. Haas (2002) then proceeds to define the *separable* spatio-temporal covariogram as a product of the spatial and temporal covariograms, and discusses some examples of the processes with long memory and/or temporally asymmetric cross-covariograms.

Clearly, this local approach is likely to be beneficial for non-stationary processes. Another way to capture non-stationarity might be to search for a non-linear transformation of the field that would make it "more stationary", in a way similar to what variance stabilizing transformations do in a univariate case (Sampson and Guttorp 1992). A semiparametric modification of this idea with the estimation of the parameters of the covariance matrix and the radial basis function representation by the maximum likelihood is discussed in Smith (1996).

The current paper takes an approach similar to Holland *et. al.* (2000). They put forward a generalized additive model that accounts for both temporal, seasonal and spatial trends, as well as controls for the meteorological variables, and model the remaining residual variability as a homogeneous Gaussian random field. They also clustered monitoring sites, although with a different purpose than in this paper, namely with that of forming the separate areas over which the inference is made.

There is a number of practical problems one faces with the contemporary spatio-temporal data, as it is often the case with the repeated measurement / longitudinal / panel data. A universal large sample problem is high power of hypothesis testing methods that leads to rejection of pretty much any hypothesis (large sample curse). Another problem is that it is extremely difficult to sustain the complete data matrix with many observation sites in long periods of time, so missing data is a common and typical problem with longitudinal data.

One of the most appealing ways to deal with the missing data, at least in the models estimated by the maximum likelihood, is to use the EM algorithm. This is an iterative algorithm that alternates between prediction of the missing data, or the sufficient statistic of it, and maximization over the parameter space. Under some minimal conditions, it converges to the maximum likelihood estimates. We shall make use of a version of this algorithm in estimating the parameters of the spatio-temporal process.

The structure of the paper is as follows. Section 2 describes the data and poses the research questions. Section 3 presents the semiparametric model that accounts for trends in space and time, as well as for the residual spatial covariance. Then Section 4 describes the principle of the EM algorithm, and shows

[2]Haas (2002) acknowledges that the standard errors are biased downwards as they do not account for the fact that the parameters of the generating distribution were estimated, and suggests using Bayesian methods to fully account for this effect.

how it can be applied in our setting. Section 5 presents the estimation results, and Section 6 concludes.

2 The data

The data used in this research is a part of the EPA data set for 1999 on the monitors of particular matter. The total number of the continental US monitors in the data set is 780. The measured variable is the concentration of the particular matter with the aerodynamic diameter of the particle less than 2.5 microns ($PM_{2.5}$), which is the policy variable regulated by the 1996 federal standard. The observation frequencies generally vary from site to site. The majority of sites have observations recorded once in three days; there are some that have daily records, and there are some that only have a few observations for the whole period. The characteristics of the monitor itself include the geographic position (latitude and longitude), the area type as a combination of two categorical factors, urbanization (rural, urban, suburban) and land use (agricultural, industrial, commercial, residential, forest)[3], altitude of the monitor, the testing method, and some other technical information.

We only used a fraction of this rich data set related to North Carolina, South Carolina, and Georgia. There were 74 monitors across those states (23 in Georgia, 31 in North Carolina, 20 in South Carolina). No data is available for Georgia in the fourth quarter of the year. The data were further aggregated into the weekly averages, so that the data set is filled more regularly. Some biases might have been introduced at this stage due to the day of the week effect. The $PM_{2.5}$ concentrations are generally lower on the weekends when there is not as much industrial activity and traffic as during the business days, so if the weekends were under- or overrepresented in a given week, then the weekly average would be biased up- or downwards.

We ended up with 2765 observations. The proportion of missing data is rather high: compare the above figure with $74 \times 52 = 3848$ observations that should be in the complete data set.

3 The spatio-temporal model

3.1 The time trend

Preliminary analysis of the data showed that the time trend can be isolated from the data, along with additive effects of the monitor location. A flexible semiparametric from of the trend was chosen, namely, the B-splines (Green and

[3]Some cells are empty: there are no combinations of forest and urban or suburban, as well as agricultural and urban

Silverman 1994). B-spline builds an approximation to a function (or a collection of observed points) as the weighted sum of the basis functions:

$$B(x) = \begin{cases} \frac{3|x|^3-6x^2+4}{6}, & -1 \le x \le 1, \\ \frac{(2-|x|)^3}{6}, & 1 < |x| \le 2, \\ 0, & 2 < |x|. \end{cases} \tag{1}$$

$$\hat{f}(t) = \alpha_0 + \sum_{k=1}^{K} \alpha_k \delta_k(t), \quad t \in [0,T], \quad \delta_k(t) = B\Big(\frac{K}{T}\big(t - Tk/K\big)\Big) \tag{2}$$

Coefficients $\alpha_0, \ldots, \alpha_K$ can be estimated by any reasonable method — say, OLS or GLS, in case the errors are (spatially) correlated. The smoothing parameter K is the number of *knots*, i.e., points to which the basis functions are tied. The choice of K, as it always happens in the smoothing world, is based on the trade-off between the bias and the variance of the functional estimate. Smaller K lead to greater smoothing, and thus may introduce the oversmoothing bias. On the other hand, splines with larger K are insufficiently smooth.

Fig. 1 shows the comparison of several fitted trends with different number of knots. The values of K are 12 (i.e., one per month, which roughly corresponds to using monthly averages), 20, or 40[4]. The graphs in the top row are those for separate states. The following five graphs show different subgroups of the monitors according to the nearby land use, and the graph in the bottom right gives the comparison of the fitted trend. Note that the splines are the same on all graphs, with the parameters estimated from all of the data (the last panel). Judging from the fit of those graphs, we have chosen the value $K = 20$ for subsequent analyses.

3.2 The spatial trend

The trend in space was also estimated non-parametrically via the bivariate version of splines, known as *thin-plate splines* (Green and Silverman 1994). The basis function for this spline evaluated at the point (x, y) is given by

$$\psi(x,y) = \frac{r \log r}{16\pi} \tag{3}$$

where $r = \sqrt{x^2 + y^2}$ is the distance from the origin (i.e., the knot of the spline). Note that unlike the unidimensional case, $\psi \to \infty$ as $(x,y) \to \infty$. Then the spatial trend can be represented as

$$\Psi(\mathbf{z}) = \beta_x x + \beta_y y + \sum_{j=1}^{J} \beta_j \psi(z_1 - x^{(j)}, z_2 - y^{(j)}) \tag{4}$$

[4] $K = 52$ would make the model a saturated one: there will be one estimated parameter for each of the weeks, and we might have used weekly averages across all monitors, instead.

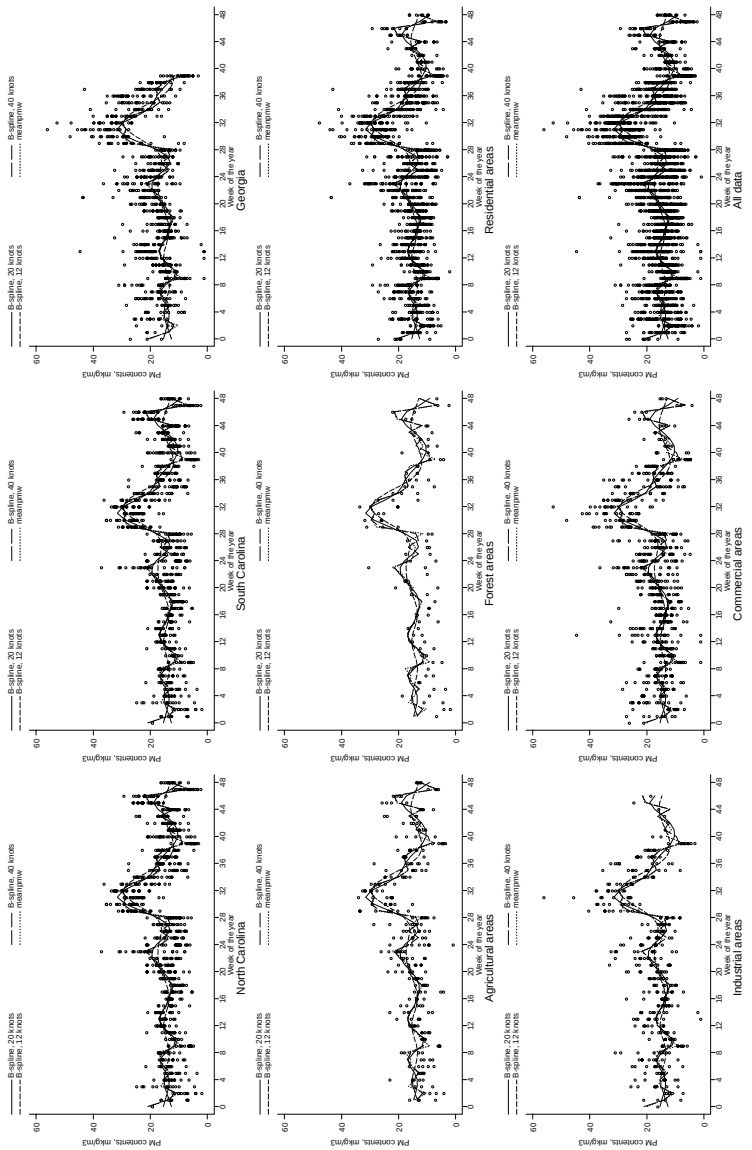

Figure 1: The comparison of the fitted trend and the raw data for subpopulations.

Again, the number of knots is the smoothing parameter. Green and Silverman (1994) operate with the case of J equal to the total number of sites (i.e., one knot per monitor). It might be desirable to reduce the number of knots to have a smoother surface. One of the options is to use a regular grid for knots of the spline. Another option that appears more flexible is to use clustering routines to place knots in the centers of the clusters of monitors, and that is the option used in our paper. We had 20 basis thinplate functions with knots in the center of clusters as found by the k-means routine. As long as we have 74 sites, the results are essentially equivalent to putting the knot in the center of three or four adjacent sites.

Another part of the model that can be thought of as a component of the spatial trend are the additive terms that account for differences in the landscape surrounding the observation site. There are 11 dummies for 12 categories — see the footnote on page 4.

3.3 Error covariance

Even after the removal of the geographical trend, some spatial variation can be expected to be present. The error process is assumed to be homogeneous in space and uncorrelated over time. The exponential power family of function was used as the functional form of the variogram:

$$\mathrm{Cov}\left(\epsilon_{ti}, \epsilon_{sj}\right) = 2\alpha(1 - \text{nugget}) \exp\left(-\left[\frac{d(site_i, site_j)}{R}\right]^p\right) \delta_{st}, \tag{5}$$

where δ_{st} is Kronecker's symbol, $\alpha > 0$ is the overall variance parameter (i.e. the variance at each particular site; that is, homoskedasticity across sites is assumed); $d(\cdot, \cdot)$ is the distance between the sites (here, the Euclidean distance between the geographic coordinates as if they were on a plane); R is the range parameter; p is the power (shape) parameter (special cases: the exponential model for $p = 1$, and Gaussian model for $p = 2$); and the nugget effect is allowed for (Smith 2000)[5]. All four parameters can be fixed or assumed flexible in the current version of the program; however, it does not seem to make much sense to have α and R fixed, as those certainly need to be estimated.

3.4 Overall model

Thus, the model that we are estimating is of the following form:

$$y_{it} = \phi_{spatial}(i) + \phi_{temporal}(t) + \phi_{individual}(i) + \varepsilon_{it}, \tag{6}$$

[5]There might have been nicer and richer parametric forms like Matérn class of functions (Smith 2000), but the currently used software (StataCorp. 2001) does not have Bessel functions implemented.

where $\phi_{spatial}(i)$ is the spatial trend modelled by thin plate splines, $\phi_{temporal}(t)$ is the temporal trend modelled by B-splines, $\phi_{individual}(i)$ is the additive term for the area type, and the error terms ε_{it} are assumed to be uncorrelated over time, but correlated over space, so that $\forall i \operatorname{Cov} \varepsilon_{it} = \Sigma$.

The parameter vector includes parameters for the spatial trend in the thin plate spline basis expansion, parameters for the temporal trend, parameters for the category of the site, and the parameters of the covariance matrix (i.e., parameters of the error variogram).

4 The EM algorithm

If the additional assumption that the errors follow a suitable multivariate normal distribution is made, then the model can be estimated by the maximum likelihood.

The fact that some of the data are missing is formally not much of a trouble for maximum likelihood. We can still write down the likelihood function for each moment in time t as

$$l(\theta|y_t) = (2\pi)^{-n_t/2}|\Sigma_t(\theta)| \exp\big[-\frac{1}{2}(y_t - x_t\beta_t)\Sigma_t(\theta)^{-1}(y_t - x_t'\beta_t)\big] \qquad (7)$$

where the subindex t indicates that observations from different sites are available at different points in time, so the dimensions of the measured $PM_{2.5}$ concentration y_t, the explanatory variables x_t, and the vector of the various trends coefficients β_t, are changing from one week to another, according to the number of available sites. Computing many determinants and the inverse matrices is likely to be time consuming, so other alternatives might be sought for. One such alternative is the EM algorithm.

The EM algorithm is an iterative algorithm that is able to obtain the ML estimates of a parametric model in the presence of missing data. The term was introduced by Dempster, Laird and Rubin (1977) where the main convergence results were also proved, and suggested monographs on the topic are Little and Rubin (1987) and McLachlan and Krishnan (1997). The algorithm delivers the estimates for the case when the data are missing at random (MAR), i.e. the probability that a given variable is not observed in a given case is independent of the "true value" of that variable that would be observed otherwise[6].

The algorithm alternates expectation (E) and maximization (M) steps. At the expectation step, the conditional expected value of the log likelihood given

[6] A more difficult case is when the missing mechanism is non-ignorable, i.e. the probability that the datum is missing depends on its true value. A simpler case is when the data are missing completely at random (MCAR): the probability that the given variable is not observed is constant across the whole sample, i.e. the probability of not observing the data conditional on all other variables is in fact an unconditional probability (Little and Rubin 1987).

the observed data Y_{obs} and the current value of the parameter vector $\theta^{(h)}$ is recomputed by using the underlying parametric model for the probability of being missing (notation follows Little and Rubin (1987)):

$$Q(\theta|\theta^{(h)}, Y_{obs}) = \int l(\theta|Y) f(Y_{miss}|Y_{obs}, \theta = \theta^{(h)})\, dY_{miss} \tag{8}$$

One can think of this step as of a sort of imputation step, although imputation usually refers to coming up with a number for a missing datum, while the E step of the EM algorithm may also deal with other moments, cross-products, etc. Another expression often used is "to integrate out" the missing values. In fact, if there is a sufficient statistic for the model, then it is enough to compute the expected value of this statistic conditional on the observed values of the variables involved, and on the current parameter values. At the maximization step, the full likelihood is maximized with respect to the parameters by using the "imputed" missing values or the expected values of the sufficient statistic:

$$\theta^{(h+1)} = \arg\max Q(\theta|\theta^{(h)}, Y_{obs}) \tag{9}$$

The procedure is iterated until convergence, which is that the successive parameter values do not change much, or the likelihood does not change much, or any sensible combination of the two.

It is proved (Little and Rubin 1987) that EM algorithms converge to stationary points of the log likelihood functions under regularity conditions that seem to be quite general (some sort of smoothness of the likelihood function, interchange of expectation and differentiation operators, boundedness of the likelihood function from above). There are also versions of the algorithm (generalized EM, or GEM) that do not maximize the likelihood at the M step, but just increase it:

$$Q(\theta^{(h+1)}|\theta^{(h)}, Y_{obs}) \geq Q(\theta^{(h)}|\theta^{(h)}, Y_{obs}) \tag{10}$$

Even in this setting, GEM algorithms converge.

One of the weak point of EM algorithms that needs to be mentioned is that they do not produce standard errors in the way Newton-Raphson likelihood maximization procedures do.

4.1 The implementation

Apparently, the missingness is concentrated on the response variable, which is the measurement of the $PM_{2.5}$ concentrations, $\mu g/m^3$. All the design variables are observed perfectly. In using the EM algorithm, we implicitly assume that the data are missing at random. This assumption would be violated if an observation is not registered when the observed value is too high or too low.

For the GEM algorithm, the maximization step was split into two steps each maximizing the likelihood over a partition of the parameter space. At the first stage, the log likelihood is maximized over the covariance matrix parameters subspace with fixed values of the additive model parameters. Then the fitted values are substituted for missing observations (or, equivalently, zero values are substituted for the residuals if the latter are missing), and the h-th step estimates of the residual (co)variances are substituted into the sufficient statistic (which in this case is the outer product of the residual vectors) if both observations are missing. Then the likelihood function is maximized over the four parameters of the covariance matrix (nugget, overall variance α, scale parameter R, shape parameter p). The whole procedure was coded in Stata software (StataCorp. 2001)[7].

Then at the second stage of the M step, a GLS regression model is estimated with the current covariance matrix estimate thus optimizing over the regression parameters subspace.

The expectation step predicts the fitted values for the GLS regression, and calculates the current step EM predictions of the Y's:

$$Y_{EM\,fit} = \begin{cases} Y_{obs}, & Y \text{ is non-missing} \\ x'\beta_{GLS}^{(h)}, & Y \text{ is missing} \end{cases} \tag{11}$$

Then the residuals e are extracted, and the process reiterated: (the conditional expectation of) the sufficient statistic ee' is calculated, where the current estimates of the covariances (i.e., the elements of the covariance matrix) are used when both residuals are initially missing, and so on.

The starting values of the parameters for the algorithm are the available case OLS regression results for the regression part of the parameter vector, and some "reasonable" guesses for the covariance part. See details in the next section.

4.2 Kriging

As is readily seen, the universal kriging formula (e.g., (2.60) of Smith (2000)) is equivalent to the following:

$$\hat{y}_0 = x_0^T\hat{\beta} + \tau^T\Sigma^{-1}e \tag{12}$$

where y_0 is the best linear prediction at the point characterized by the regressors x_0; $\hat{\beta}$ is the estimate of the regression coefficient of the process $Y = X\beta + \nu$, $\operatorname{Cov}\nu = \Sigma$, so that $x_0^T\hat{\beta}$ is the linear fit, or the trend term, from the model; τ is the vector of cross-covariances between the observed values of the field Y

[7]The description of the maximum likelihood procedure built into the package is given in the Appendix.

and the unobserved y_0 given by the spatial model; and e is the residuals of the process from the fitted linear regression: $e = Y - X\hat{\beta}$.

The linear fit term arises naturally from the EM algorithm. This is the predicted value $Y_{EM\,fit}$ computed at the E step. In fact, if one introduces completely missing observation into the data set, one can get predictions for it along the way, but this would be rather inefficient from the computational point of view, as the pace of convergence of EM algorithm depends on the ratio of missing information to the complete information, and thus it is not recommended to introduce fully missing variables unless it is absolutely necessary.

To implement kriging following the ML (or EM) estimation, the estimates $\hat{\beta}, \hat{\Sigma}$ obtained at the last iteration can be used.

5 Results

Several runs of the algorithm were tried with different pseudorandom number generator seed that were supposed to produce different clustering of the monitors, and thus slightly different models for the spatial trend. The covariance matrix estimates when fully flexible model is used are as follows: $\alpha = 2.917$ (reported s.e. 0.049), $p = 1.444$ (reported s.e. 0.139), $R = 2.305$ (reported s.e. 0.210)[8], and nugget $= 0.450$ (reported s.e. 0.027). The reported s.e.s are naive as they are obtained with the full data matrix assumption.

(The following is a very *ad hoc* argument.) A crude way to obtain an idea of how large the standard errors should be is to decompose the information matrix into the observed and missing information. Section 7.5 of Little and Rubin (1987) derives the expectation of missing information due to missing data and gives a simple interpretation of the resulting matrix expression that

$$\text{observed information} = \text{complete information} - \text{missing information}.$$

Assuming a simple proportionality, observed information = proportion of observed cases $\times$ complete information. Hence, the information is overestimated by the factor of $2613/3626 \approx 0.721$ where the numerator is the number of observed cases and the denominator is the total number of cases (74 sites $\times$ 49 weeks). Thus, the standard errors should be inflated by some 18% to be a better approximation of reality. As long as a great proportion of the missing data is accumulated in Georgia in the last quarter of 1999, it should be expected that the correct standard errors should be adjusted relatively heavier for the last 5

[8] This corresponds to the distance of about 168.9 miles, s.e. 24.6 miles. At this distance, the estimate of correlation is about 0.2. Li *et. al.* (1999) report the correlation scale of the PM_{10} hourly measurements of about 7.5 km. In a later paper (Li *et. al.* 2000), they use Sampson and Guttorp (1992) approach to non-parametrically estimate the spatial covariance matrix, and note that the field is not isotropic.

Table 1: Area type effects.

Area type	Rural	Suburban	Urban
Agricultural	-1.716	-1.585	n/a
# sites	5	1	0
# obs.	186	44	0
Commercial	1.162	0.935	-0.267
# sites	1	10	8
# obs.	24	348	272
Forest	-1.667	n/a	n/a
# sites	3	0	0
# obs.	101	0	0
Industrial	-0.943	-0.692	-0.805
# sites	1	7	2
#obs	32	251	78
Residential	-2.647	base	0.021
# sites	1	18	17
# obs.	41	652	584

or so terms in the temporal trend, and for the sites in Georgia in the spatial trend.

The area type $\phi_{individual}(i)$ term was described by a set of categorical dummies. The estimated contrasts are given in the Table 1.

The results of kriging are presented at Fig. 2–6. The base (prevalent) category "suburban-residential" was selected for the area type term $\phi_{individual}(i)$ of (6). The estimates of the dummies in Table 1 should be added on top of those kriging estimates for different landscape / land use categories. The patterns of $PM_{2.5}$ contents vary over the year, but the general pattern is that the level of $PM_{2.5}$ is higher in Georgia than in Carolinas, and tends to decrease towards the coast and further into the ocean. This seems to support the point of view that the $PM_{2.5}$ is by and large a by-product of human activities. As the earlier trend analysis had shown, there is a spike of $PM_{2.5}$ levels in late summer, which is seen at Fig. 5.

6 Conclusions

This paper proposed and exemplified the use of likelihood based methods in the generalized additive model framework, with trends accounting for (most of the) variation in space and time, as well as across the sites in different area types.

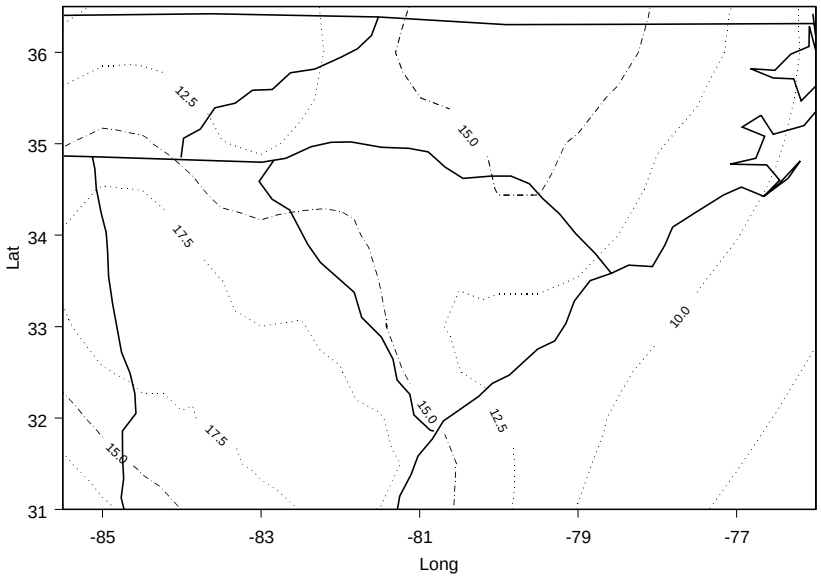

Figure 2: Week 1 of observations.

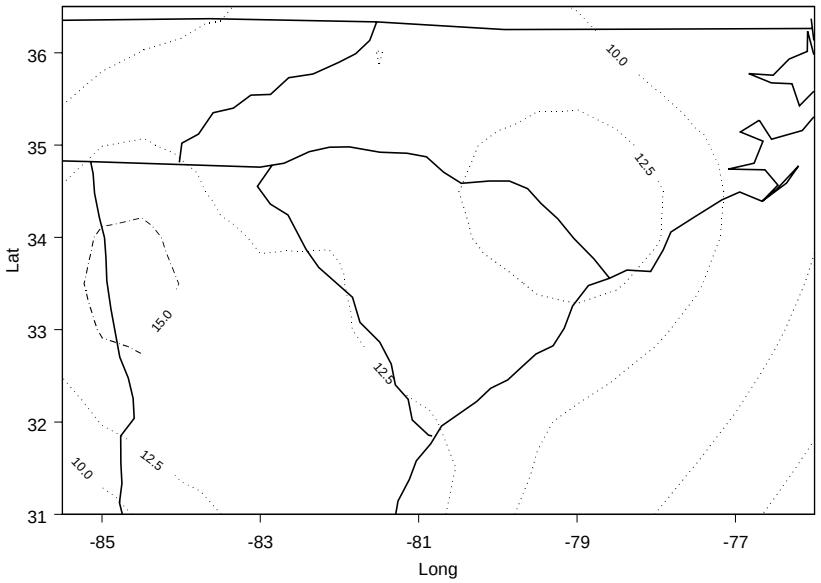

Figure 3: Week 10 of observations.

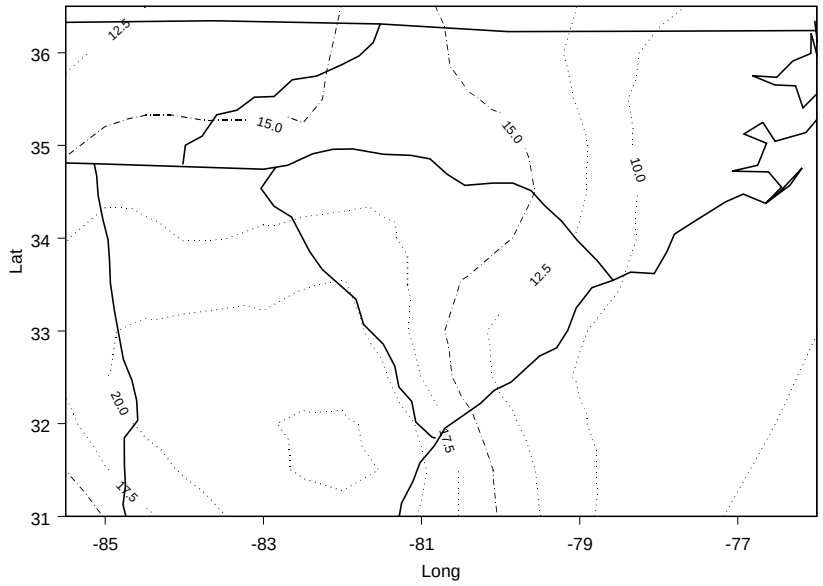

Figure 4: Week 20 of observations.

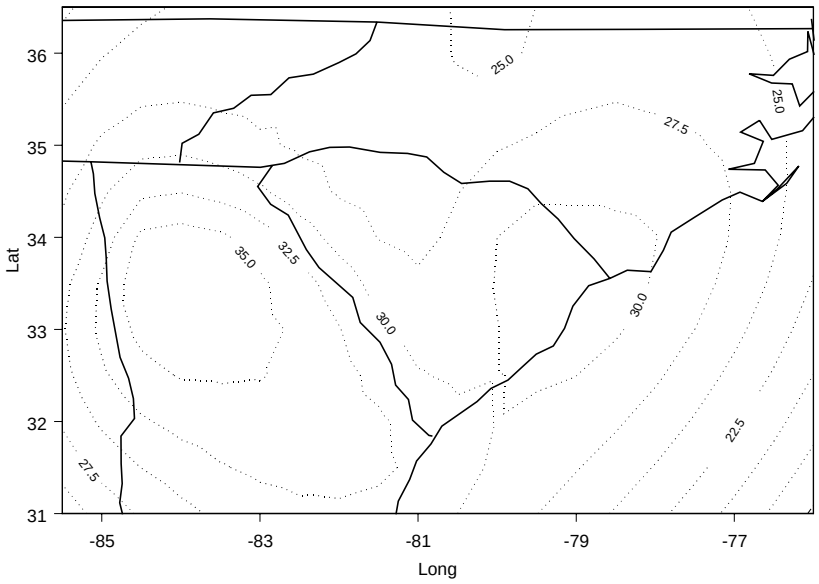

Figure 5: Week 30 of observations. Note that all $PM_{2.5}$ levels are higher than 15 $\mu g/m^3$.

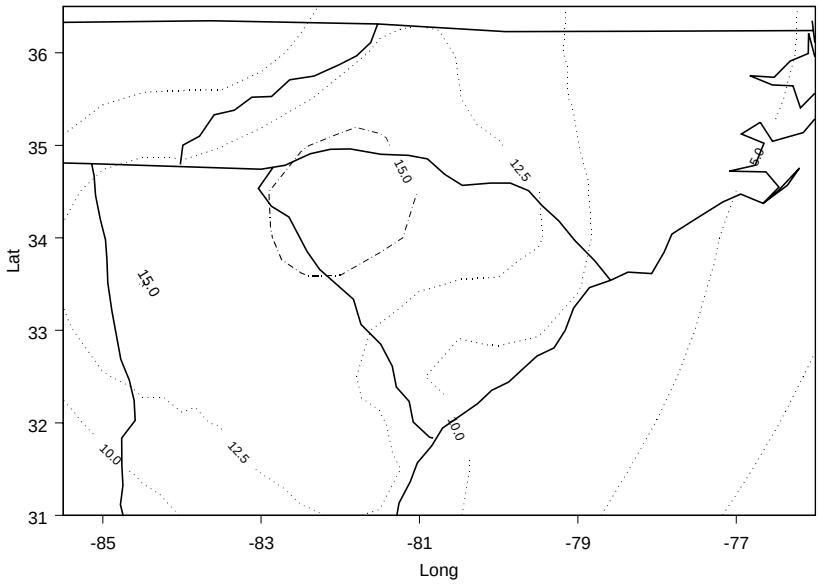

Figure 6: Week 40 of observations.

The estimation of the parameters is done through the EM-algorithm to correct for missing data in the longitudinal data sets. The procedure helped to retain a large fraction of monitors even though they reported the data infrequently. The estimates of the spatial covariance matrix suggest that the range of the error correlations is about two hundred miles, and the shape of the dependence on the distance between two sites is between the exponential and the Gaussian decay.

The substantive results imply that the three analyzed states (Carolinas and Georgia) are in danger of violating the federal standard on $PM_{2.5}$, except for the coastal areas, Appalachians, or during the winter months. (This statement might be attenuated by the fact that the monitors might have been located in the "problematic" areas that are known to have polluted air, so that the design of the monitoring network is not a random one.)

Acknowledgements

We are grateful to Dave Holland who kindly provided the data to us; to Ken Bollen, for very fruitful discussions of the approaches to the missing data analysis; and to the participants of the meetings where this research was presented. The research was financed through the EPA grant.

Appendices

A Stata software

Stata software (StataCorp. 2001, Kolenikov 2001) is a general purpose statistical package that features a built-in procedure to perform the maximum likelihood estimation of a function specified by the user (Gould and Sribney 1999). The main steps of the ML algorithm are the following.

1. Optionally, the ML code can be checked for blunt errors, such as no likelihood is calculated for any values; different values of the likelihood are returned for the same parameter values; and some other consistency checks.

2. Optionally, user can specify some starting values for the algorithm.

3. By default, Stata searches randomly for the initial values of the parameters over the parameter space. It also reverts to a random search if the initial values specified by the user are not feasible (the log likelihood cannot be calculated). User can specify the search bounds for particular values of parameters or equations.

4. By default, Stata performs some simple attempts to improve the initial values by scaling each of the parameters or equations. As a rule, the ML algorithm finds pretty good estimates by now.

5. Finally, Stata starts multidimensional optimization of the parameter vector. In the simplest case, the user only codes the likelihood of the i.i.d. observations. This is not applicable for our case, however, as long as there are (spatial) dependencies across sites, so I used another option: coding the value of the likelihood function to be provided to the optimizer for any combination of the parameter values (may be a missing value if the likelihood cannot be calculated, e.g. if the range parameter is negative). Other advanced options of coding the likelihood include possibilities to specify the gradient and Hessian matrix of the likelihood; if those are not provided by the user, Stata uses numerical derivatives as described in Gould and Sribney (1999).

6. The procedure is terminated when the relative change in the parameters between the two consecutive iterations is small, or absolute change in the objective function is small, or a maximum number of iterations is performed. In addition, another necessary criterion can be specified that the norm of the gradient vector is sufficiently small.

While performing the multidimensional maximization, the ML optimizer is tracking concavity of the likelihood function and produces warning messages if the likelihood function is not concave. If an option of doing more work for such occasion is specified, then the ML algorithm solves an eigenproblem for the current value of the Hessian matrix, and combines Newton-Raphson steps along the directions in the subspace where the eigenvalues are found to be positive (up to numerical accuracy, defined in a special way) with the steepest accent in the subspace of negative, zero, or small positive eigenvalues. It is an evidence of lack of convergence if "not concave" message is produced at the last iteration. In this application, the lack of convexity was not found to be a problem, so this option has never been specified.

It turned out that the operation taking most time was filling in the current estimate of the covariance matrix from the given parameters that involved approx. $(\#\text{sites})^2$ "slow" operations. Each call need to have been interpreted and parsed by Stata, while operations like matrix inversion and determinant calculation are internal to Stata core and thus are relatively fast.

References

Cox, L. H. (2000). Statistical issues in the study of air pollution involving airborne particular matter. *Environmetrics*, **11**, 611–626.

Cressie, N. (1993). *Statistics for Spatial Data*. 2nd edition. John Wiley, New York.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *JRSS*, **B39**, 1–38.

Green, P. J., and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Monographs on Statistics and Applied Probability, Vol. 58. Chapman & Hall.

Gould, W., and Sribney, W. (1999). *Maximum Likelihood Estimation with Stata*. Stata Press, College Station, TX.

Haas, T. (1998). Statistical assessment of spatio-temporal pollutant trends and meteorological transport models. *Atmos. Environment*, **32**, 1865–1879.

Haas, T. (2002). New systems for modeling, estimating, and predicting a multivariate spatio-temporal process. *Environmetrics*, in press.

Holland, D. M., De Olivera, V., Cox, L. H., and Smith, R. L. Estimation of regional trends in sulfur dioxide over the eastern United States. *Environmetrics*, **11**, 373–393.

Kolenikov, S. (2001). Review of Stata 7. *J. of Applied Econometrics*, **16**, 637-646.

Li, K. H., Le, N. D., Sun, L., and Zidek, J. V. (1999). Spatial-temporal models for ambient hourly PM_{10} in Vancouver. *Environmetrics*, **10**, 321–338.

Li, K. H., Zidek, J. V., Le, N. D., and Ozkaynak, H. (2000). Interpolating Vancouver's daily ambient PM_{10} field. *Environmetrics*, **11**, 651–663.

Little, R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics, John Wiley, New York.

McLachlan, G.J., and Krishnan, T. (1997) *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics, John Wiley, New York.

Sampson, P., and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance matrix. *JASA*, **87**, 108–119.

Smith, R. L. (1996). Estimating nonstationary spatial correlations. Unpublished manuscript, University of North Carolina, Chapel Hill.

Smith, R.L. (2000). *Environmental Statistics*. Lecture notes, http://www.stat.unc.edu/postscript/rs/envnotes.ps.

StataCorp. (2001). *Stata Statistical Software: Release 7*. College Station, TX: Stata Corporation.