Bivariate Splines for Ozone Concentration Forecasting

Bree Ettinger
Department of Mathematics
The University of Georgia
Athens, GA 30602, USA

Serge Guillas*
Department of Statistical Science
University College London
UK

Ming-Jun Lai†
Department of Mathematics
The University of Georgia
Athens, GA 30602, USA

Abstract

In this paper, we forecast ground level ozone concentrations over the USA, using past spatially distributed measurements and the functional linear regression model. We employ bivariate splines defined over triangulations of the relevant region of the USA to implement this functional data approach in which random surfaces represent ozone concentrations. We compare the least squares method with penalty to the principal components regression approach. Moderate sample sizes provide good quality forecasts in both cases with little computational effort. We also illustrate the variability of forecasts due to the choice of smoothing penalty. Finally, we compare our predictions with the ones obtained using thin-plate splines. Predictions based on bivariate splines require less computational time than the ones based on thin-plate splines and are more accurate. We also quantify the variability in the predictions arising from the variability in the sample using the jackknife, and report that predictions based on bivariate splines are more robust than the ones based on thin-plate splines.

Keywords: functional data, ozone, bivariate splines.

1   Introduction

Ground-level ozone is a harmful pollutant (Bell et al., 2004). To inform the public on a daily basis about ozone and other pollutants, the US Environmental Protection Agency (EPA) has created a simplified index of air quality, the pollutant standards index (PSI), and since July 1999, EPA replaced the PSI with the Air Quality Index (AQI). Forecasts of these indices made 24 hours ahead have been provided to newspapers in various regions so that people can avoid outdoor activities likely to damage their health. Using the PSI, Neidell (2010) showed that air quality warnings associated with ground-level ozone have had an impact on outdoor activities in Southern California, especially for susceptible local residents. Hence, improving the quality of these forecasts may contribute to better public health.

We consider here ozone concentrations from the network of EPA stations across the USA, over a span of three months in the summer of 2005. Our goal is to predict the ozone concentration values at a specific location in suburban Atlanta 24 hours ahead based on the previous ozone concentrations values up to the current hour. There are many available methods, ranging from chemistry-transport models to statistical techniques. Recently, Guillas and Lai (2010) initiated a brand new statistical approach to predict ozone concentration values using functional linear regression with bivariate splines (piecewise polynomial functions over a triangulation of a polygonal domain). The goal is to capture predictive power in time from spatial synoptic scales, as chemistry and transport occur at regional scales. Such spatio-temporal information has been used before to model time-dependent dynamics in hourly ozone concentration fields, see e.g. Dou et al. (2010). However, our approach represents the surface as a random function, which enables us to carry out regression and prediction in that setting. Modelling observations as functional data presents many advantages, see Ramsay and Silverman (2005) and Ferraty and Vieu (2006) for an overview of functional data analysis (FDA). For that purpose, one needs a parametric representation of the surfaces ideally in a flexible basis that kriging is not able to offer. Guillas and Lai (2010) established the theoretical foundations, even under time-dependence assumptions, of this method for which a least squares criterion with penalty is minimized. In the following, we refer to this approach as the "brute force method" (BF). Comparison with time series approaches using FDA (i.e. in one dimension: time) shows that BF can predict better, particularly with small sample sizes by borrowing strength across space around the location of interest, see Guillas and Lai (2010) for examples.

Crambes et al. (2009) minimized a residual sum of squares subject to a smoothness penalty as in BF, but modified the usual penalty term of univariate smoothing splines to study the prediction.

They showed that the rates of convergence of the estimates of the slope are optimal. Yuan and Cai (2010) demonstrated that BF can deliver optimal rates of convergence for either prediction or estimation. Their setting unifies the treatment of estimation and prediction under the umbrella of one family of norms. They showed the interesting property that the faster the decay in the eigenvalues of the covariance operator of the explanatory variable, the faster the prediction and the slower the estimation of the slope.

The Functional Principal Component Regression (FPCR) has been introduced in two papers (Cardot et al., 1999, 2003), which cover the case of random functional data that are curves. It relies on the principal component decompositions prior to inference about the regression slope. Cardot et al. (2003) used univariate splines to approximate the empirical estimator for the regressor function associated with the random functional. FPCR was shown to attain optimal rates of convergence for prediction and estimation (Cai and Hall, 2006; Hall and Horowitz, 2007). As pointed out by Cressie and Wikle (2011), the Empirical Orthogonal Functions (EOFs) employed in space-time analysis of geophysical data are merely principal components for the vector of observations. In practice, EOFs are used since observations are collected at discrete points. However, this approach can yield misleading results as these observations are not weighted spatially. The analysis should reflect the continous variations of the field. The Karhunen-Loève expansion considered below is the continuous equivalent of the EOF decomposition. Using bivariate splines, we respect the continuous nature of ozone fields.

In our context of ozone forecasting, Aneiros-Perez et al. (2004) considered a functional additive model and Crambes et al. (2009) applied least squares with penalty to the prediction of the maximum of the ground-level ozone concentration. However, these studies do not exploit spatial information. When dealing with random surfaces and a functional associated with these over a domain of irregular shape, bivariate splines can be an excellent approximation tool, especially when the random surfaces are observed only over scattered locations (Lai, 2008) as these splines possess optimal approximation properties (Lai and Schumaker, 2007). Thus, we shall use these splines in FPCR and would like to know how it compares with BF for the ozone concentration forecasting problem. We shall explain this generalization in the following section in detail and discuss the similarities and differences between BF and FPCR in this context. We illustrate numerically the advantages and disadvantages of these two approaches. The biggest difference lies in the fact that BF is straightforward and can be used for prediction without a spectral decomposition while the FPCR requires the use of a selected number of eigenvalues and eigenvectors in advance to be effective for prediction.

Thin-plate splines have been used in various applications to model surfaces, see Wood (2003) and references therein. For instance, Paciorek et al. (2009) considered the measurement of particulate matters concentrations, and used thin-plate splines to represent spatial variations in a spatio-temporal model. We wish to compare the skills of thin-plate splines and bivariate splines in our problem of spatial representation and prediction of ozone concentrations. For that aim, we compare in this paper BF using either thin-plate splines or bivariate splines. This is the first time that such comparison is carried out.

The paper is organized as follows. We first explain two approaches of prediction with an emphasis on the FPCR in the next section and we explain the similarities and differences of these two methods in practice. Then we present numerical experiments in Section 3. In Section 4, we compare the use of thin-plate splines and bivariate splines. In section 6, we address the issue of the choice of triangulation. Finally in section 7, we conclude and discuss some future research directions.

2 Two Approaches of Forecasting

Let Y be a real-valued random variable which is a functional of random surface X. That is, let $\mathcal{D}$ be a polygonal domain in $\mathbf{R}^2$. The functional linear model for Y is:

$$Y = f(X) + \varepsilon = \langle g, X \rangle + \varepsilon = \int_{\mathcal{D}} g(s)X(s)ds + \varepsilon, \tag{1}$$

where X is a random surface over $\mathcal{D}$, $g(s)$ is in a function space H (usually $L^2(\mathcal{D})$), ε is a real random variable that satisfies $E\varepsilon = 0$ and $EX(s)\varepsilon = 0, \forall s \in \mathcal{D}$.

2.1 Brute-Force (BF)

The BF method is explained in detail in Guillas and Lai (2010). For convenience, let us outline the approach as follows. The estimate of the function $g \in H$ is chosen to solve the following minimization problem:

$$\min_{\beta \in H} E\left[(Y - \langle \beta, X \rangle)^2 \right] + \rho\|\beta\|_2^2, \tag{2}$$

where $\rho > 0$ is a penalty, and $\|\beta\|_2^2$ denotes the semi-norm of β:

$$\|\beta\|_2^2 = \int_{\mathcal{D}} \sum_{i+j=2} \left(D_1^i D_2^j \beta \right)^2, \tag{3}$$

in which D_k denotes the differentiation operator with respect to the k-th variable. Our objective is to determine or approximate g which is defined on a spatial domain $\mathcal{D}$ based on observations of X, from a set of design points in $\mathcal{D}$, and of the random variable Y. We choose a spline space $S_d^r(\triangle)$, i.e. the space of polynomials of degree d and smoothness r over the triangulation $\triangle$ (Lai and Schumaker, 2007), which is a finite dimensional subspace of $L^2(\mathcal{D})$.

In practice, the random surfaces X_i's are not observed continuously but at design points $s_k \in \mathcal{D}, k = 1, \cdots, N$. The smooth bivariate spline approximation S_{X_i} of X_i is the solution in $S_d^r(\triangle)$ of the following minimization

$$\min \left\{ \sum_{k=1}^{N} |h(s_k) - X_i(s_k)|^2 + \gamma \mathrm{En}(h), h \in S_d^r(\triangle) \right\},$$

where $\mathrm{En}(h) = \int_{\mathcal{D}} (D_1^2 h)^2 + 2(D_1 D_2 h)^2 + (D_2^2 h)^2$ is a energy functional and $\gamma > 0$ a smoothing parameter. The existence, uniqueness and computational scheme can be found in Awanou et al. (2006). The approximation properties of the penalized least squares fit are summarized in Lai (2008).

We seek an approximation $\widetilde{S_{g,n}} \in S_d^r(\triangle)$ of the empirical estimator of g such that $\widetilde{S_{g,n}}$ minimizes the following:

$$\min_{\beta \in S_d^r(\triangle)} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle \beta, S_{X_i} \rangle)^2 + \rho\|\beta\|_2^2 \tag{4}$$

where $S_{X_i} \in S_d^r(\triangle)$ is a penalized least squares fit of the ozone data X_i on the hour i using the bivariate spline space $S_d^r(\triangle)$. The space of bivariate splines over the domain $\mathcal{D}$, $S_d^r(\triangle)$, can be chosen for instance to be $S_1^0(\triangle)$ (linear finite elements), $S_5^0(\triangle)$ for continuity only but with polynomials of degree 5, or $S_5^1(\triangle)$ (our actual choice in this paper) for more smoothness across the domain. We choose $d = 5$ and $r = 1$ because these values satisfy the minimal requirements for a

smooth spline to have the optimal approximation order with d greater than $3r + 1$, see Th 2.1 in Guillas and Lai (2010). We do not consider here higher degrees because the additional number of degrees of freedom would require more data to yield some benefit.

Although spline functions are only C^1 differentiable in the case of $S_5^1(\triangle)$ over the domain, they are in the Sobolev space $H^2(\triangle)$ and hence, we can use the above penalty function with second order derivatives. Indeed, we compute all second order derivatives of spline functions inside each triangle (polynomials within each triangle) but not over edges nor at vertices. All edges and vertices of a triangulation form a set of measure zero for the integration in (3).

We use here the representation of splines by their coefficients in the Bernstein-Bezier basis of bivariate polynomials $\phi_i, i = 1, \cdots, m$, for which computations are efficient and continuity and smoothness conditions can easily be derived (Lai and Schumaker, 2007, Chapter 2).

The solution of the above minimization is in $S_d^r(\triangle)$ and is given by $\widetilde{S_{g,n}} = \sum_{i=1}^m \widetilde{c_{n,i}} \phi_i$ with coefficient vector $\widetilde{\mathbf{c}_n} = (\widetilde{c_{n,i}}, i = 1, \cdots, m)$ satisfying $\widetilde{A_n}\widetilde{\mathbf{c}_n} = \widetilde{b}_n$, where

$$\widetilde{A_n} = \left[\frac{1}{n} \sum_{\ell=1}^n \langle \phi_i, S_{X_\ell} \rangle \langle \phi_j, S_{X_\ell} \rangle + \rho \mathcal{E}_2(\phi_i, \phi_j) \right]_{i,j=1,\cdots,m},$$

where $\mathcal{E}_2(\alpha, \beta) = \int_{\mathcal{D}} \sum_{i+j=2} D_1^i D_2^j \alpha D_1^i D_2^j \beta$, corresponds to the semi-norm $\|\beta\|_2^2$ above, and

$$\widetilde{b}_n = \left[\frac{1}{n} \sum_{\ell=1}^n Y_\ell \langle \phi_j, S_{X_\ell} \rangle \right]_{j=1,\cdots,m}.$$

In theory, the spline space $S_d^r(\triangle)$ is dense in H as the size $|\triangle|$ of triangulation $\triangle$ decreases to zero and hence, the approximation S_g of the empirical estimator of g in (4) approximates g as the sample of n observations collected at the same hour of the day, over n consecutive days increases and $|\triangle| \to 0$ (Guillas and Lai, 2010).

2.2 Functional Principal Component Regression (FPCR)

We next spend some effort to explain the FPCR. Cardot et al. (2003) used univariate splines to approximate the function g in the functional linear model and used the principal component analysis and smoothing spline techniques to find the spline-based estimators. In the following, we generalize the ideas of Cardot et al. (2003) to deal with random surfaces over a 2D domain of irregular shape based on bivariate spline functions. Let Γ be the standard covariance operator of the H-valued random variables X, $\Gamma := E(X(s)X(t))$ and

$$(\Gamma g)(t) = \int_{s \in \mathcal{D}} E(X(s)X(t))g(s)ds, \quad \forall g \in H.$$

Let Δ be th cross-covariance of (X, Y), i.e., $\Delta := E(X(s)Y)$ with

$$\langle \Delta, f \rangle = \int_{t \in \mathcal{D}} E(X(t)Y)f(t)dt \quad \forall f \in H. \tag{5}$$

We can easily prove the relationship $\Gamma g = \Delta$.

Clearly Γ is an integral operator mapping H to H. Assume that Γ is a compact operator, as in Cardot et al. (2003). Let $\lambda_j, j = 1, 2, \cdots$, be the eigenvalues of Γ arranged in the decreasing order and $v_j \in H$ be eigenfunctions of Γ associated with λ_j for $j = 1, 2, \cdots$. Suppose that

$v_j, j = 1, 2, \cdots$, form a complete orthonormal basis for H. Then we can write $\Gamma = \sum_j \lambda_j v_j(t) v_j(s)$ and $g = \sum_j \langle g, v_j \rangle v_j$ for any $g \in H$. Hence, since Γ is a symmetric operator, we have

$$\lambda_j \langle g, v_j \rangle = \langle g, \lambda_j v_j \rangle = \langle g, \Gamma v_j \rangle = \langle \Gamma g, v_j \rangle = \langle \Delta, v_j \rangle = \langle E(X(t)Y), v_j \rangle.$$

It follows that $\langle g, v_j \rangle = \langle E(X(\cdot)Y), v_j \rangle / \lambda_j$ if $\lambda_j > 0$. Thus, we get the expansion for g:

$$g(\cdot) = \sum_{j=1}^{\infty} \frac{\langle E(X(\cdot)Y), v_j \rangle}{\lambda_j} v_j.$$

Note that the function g is in H if and only if

$$\sum_{j=1}^{\infty} \left(\frac{\langle E(X(\cdot)Y), v_j \rangle}{\lambda_j} \right)^2 < +\infty.$$

In general, we do not know if Γ is invertible or not. Let $\mathcal{N}(\Gamma)$ be the kernel of Γ. That is, $\mathcal{N}(\Gamma) = \{x \in H, \Gamma x = 0\}$ and suppose that $\mathcal{N}(\Gamma) \neq \emptyset$. Then g cannot be uniquely determined. Nevertheless, g can be determined in $\mathcal{N}(\Gamma)^{\perp}$.

Let $H_k = \text{span}\{v_1, \cdots, v_k\} \subset \mathcal{N}(\Gamma)^{\perp}$ be a finite dimensional approximation of the orthogonal complement of $\mathcal{N}(\Gamma)$. For example, we can use the spline space $S_d^{-1}(\triangle)$ of piecewise polynomial functions without smoothness over a triangulation $\triangle$ of the underlying domain $\mathcal{D}$. The discontinuous spline space $S_d^{-1}(\triangle)$ will better approximate the orthogonal complement than a spline space $S_d^r(\triangle)$ with smoothness $r \geq 1$. Next let $\mathcal{P}_k$ be the orthogonal projection operator from H to H_k. When $\lambda_k > 0$, $\mathcal{P}_k \Gamma \mathcal{P}_k$ is invertible. Note that $\mathcal{P}_k \Gamma \mathcal{P}_k g = \sum_{j=1}^{k} \lambda_j \langle v_j, g \rangle v_j$. Thus, for all $x \in H$, $P_k x = \sum_{j=1}^{k} \langle x, v_j \rangle v_j$. We have $\langle \mathcal{P}_k \Gamma \mathcal{P}_k g, P_k x \rangle = \langle \Delta, P_k x \rangle$, or

$$\sum_{j=1}^{k} \lambda_j \langle v_j, g \rangle \langle v_j, x \rangle = \sum_{j=1}^{k} \langle x, v_j \rangle \langle \Delta, v_j \rangle$$

for all $x \in H$. It follows that $\langle v_j, g \rangle = \frac{1}{\lambda_j} \langle \Delta, v_j \rangle$ for $j = 1, \cdots, k$. Hence, we obtain the approximation of g in H_k:

$$g_k = \sum_{j=1}^{k} \frac{1}{\lambda_j} \langle \Delta, v_j \rangle v_j.$$

In order to compute our estimate of g, we make use of random samples $X_i, i = 1, \cdots, n$ in H with dependent variable Y_i. We start with the case of fully observed surfaces X_i. Let Γ_n be the empirical estimator of Γ:

$$\Gamma_n x = \frac{1}{n} \sum_{i=1}^{n} \langle X_i, x \rangle X_i$$

and Δ_n be the empirical estimator of Δ:

$$\Delta_n x = \frac{1}{n} \sum_{i=1}^{n} \langle X_i, x \rangle Y_i.$$

Then the finite dimensional operator Γ_n is a compact operator mapping H to H and hence, Γ_n can be expanded in terms of its eigenfunctions $\hat{v}_j$, $j = 1, 2, \cdots$. That is,

$$\Gamma_n x = \sum_{j=1}^{\infty} \hat{\lambda}_j \langle \hat{v}_j, x \rangle \hat{v}_j.$$

Similar to the above theoretical discussion, we have

$$\Delta_n x = \langle g_n, \Gamma_n x \rangle$$

for some g_n in H. Assume that the first k_n largest eigenvalues $\hat{\lambda}_j, j = 1, \cdots, k_n$ are nonzero. Then the principal component regression estimator of g_k is obtained in H_{k_n}, the finite dimensional space spanned by $\hat{v}_1, \cdots, \hat{v}_{k_n}$:

$$\hat{g}_{PCR} = \sum_{j=1}^{k_n} \frac{\Delta_n(\hat{v}_j)}{\hat{\lambda}_j} \hat{v}_j$$

which is an approximation of the empirical estimator of g.

As we use the discontinuous spline space $S_d^{-1}(\triangle)$ to represent eigenfunctions $\hat{v}_j$, $\hat{g}_{PCR}$ is a discontinuous piecewise polynomial function, i.e., it is not continuous at the edges and vertices of $\triangle$, $\hat{g}_{PCR}$ is not in $S_d^r(\triangle)$. However, we can smooth $\hat{g}_{PCR}$ by approximating it using bivariate splines in $S_d^r(\triangle)$ with $r \geq 1$. Let $\widetilde{g}_{SPCR}$ be the solution of the following continuous least squares minimization:

$$\widetilde{g}_{SPCR} = \min_{f \in S_d^r(\triangle)} \int_{\mathcal{D}} |\hat{g}_{PCR}(s) - f(s)|^2 ds.$$

When the random sample is not fully observed, as in BF we use spline approximations of the random samples $X_i, i = 1, \cdots, n$, with penalty γ. We then use the discontinuous spline space. Let $\widetilde{\Gamma_n}$ be an approximation of the empirical estimator Γ_n of Γ:

$$\widetilde{\Gamma_n}(x) = \frac{1}{n} \sum_{i=1}^{n} \langle S_{X_i}, x \rangle S_{X_i} \tag{6}$$

and $\widetilde{\Delta_n}$ be an approximation of the empirical estimator Δ_n of Δ:

$$\widetilde{\Delta_n}(x) = \frac{1}{n} \sum_{i=1}^{n} \langle S_{X_i}, x \rangle Y_i. \tag{7}$$

Clearly, $\widetilde{\Gamma_n}$ is a bounded operator on the space spanned by bivariate polynomials, and thus we can express in the following format:

$$\widetilde{\Gamma_n}(x) = \sum_{j=1}^{m} \widetilde{\lambda}_j \langle \widetilde{v}_j, x \rangle \widetilde{v}_j, \tag{8}$$

where $\widetilde{\lambda}_j$ and $\widetilde{v}_j$ are a pair of eigenvalue and eigenvector of $\widetilde{\Gamma_n}$ and m is the dimension of the spline space $S_d^r(\triangle)$. It then follows that

$$\langle \widetilde{\Delta_n}, x \rangle = \langle g_n, \widetilde{\Gamma_n} x \rangle \tag{9}$$

for some $\alpha_n \in H$. Assume that the first k_n largest eigenvalues $\widetilde{\lambda}_j, j = 1, \cdots, k_n$ are nonzero. Then the principal component regression estimator of g_n is

$$\widetilde{g}_{PCR} = \sum_{j=1}^{k_n} \frac{\Delta_n(\widetilde{v}_j)}{\widetilde{\lambda}_j} \widetilde{v}_j \tag{10}$$

which is an approximation of the empirical estimator of g. Finally we can use the discrete least squares minimization to compute a smooth version of $\widetilde{g}_{PCR}$ and denote it by $\widetilde{g}_{SPCR}$. It is the solution in $S_d^r(\triangle)$ of the following minimization

$$\min \left\{ \sum_{k=1}^{N} |h(s_k) - g_{PCR}(s_k)|^2, \quad h \in S_d^r(\triangle) \right\},$$

where the locations of the stations are denoted s_k.

We now compare BF and FPCR. In the FPCR, the spline approximation of the covariance operator $\widetilde{\Gamma}_n$ is

$$\widetilde{\Gamma_n}(x) = \frac{1}{n} \sum_{\ell=1}^{n} \langle S_{X_\ell}, x \rangle S_{X_\ell} \tag{11}$$

from (6). For any $x = \sum_{i=1}^{m} c_i \phi_i$, the operator $\widetilde{\Gamma}_n$ maps any bivariate polynomial x into the space of bivariate polynomials as

$$\widetilde{\Gamma_n}(x) = \sum_{i=1}^{n} c_i \frac{1}{n} \sum_{\ell=1}^{n} \langle S_{X_\ell}, \phi_i \rangle S_{X_\ell}.$$

Thus the matrix associated with the covariance operator in this finite dimensional space is

$$\left[\frac{1}{n} \sum_{\ell=1}^{n} \langle S_{X_\ell}, \phi_i \rangle \langle S_{X_\ell}, \phi_j \rangle \right]_{1 \leq i,j \leq m},$$

which is the matrix $\widetilde{A}_n$ used in BF when no penalty is present (Guillas and Lai, 2010). Similarly, the spline approximation $\widetilde{\Delta}_n$ of the empirical estimator Δ_n of Δ is,

$$\widetilde{\Delta_n}(x) = \sum_{i=1}^{m} c_i \frac{1}{n} \sum_{\ell=1}^{n} \langle S_{X_\ell}, \phi_i \rangle Y_\ell, \tag{12}$$

The vector representation of $\widetilde{\Delta}_n$ is

$$\left[\frac{1}{n} \sum_{\ell=1}^{n} \langle S_{X_\ell}, \phi_i \rangle Y_i \right]_{1 \leq i \leq n}$$

which is the $\widetilde{b}_n$ used in BF.

If we were to be able to use all eigenvalues and eigenvectors of the covariance matrix $\widetilde{A}_n$ and invert the covariance operator, the solution would be the same one as the brute-force approach with no penalty. However, the FPCR approach uses a few principal eigenelements to compute g_{PCR}. Obviously, BF and FPCR can differ greatly.

As the empirical covariance matrix $\widetilde{A}_n$ is not invertible in general, the choice of regularization procedure for this ill-posed problem is either done through the addition of a penalty (as in BF), also known as the Tikhonov regularization, or by a projection on a few principal components (as in FPCR). Tuning parameters are either the penalty ρ for BF or the number of eigenvalues k_n for FPCR. For a discussion of previous work in these two approaches in the one dimension setting, see Yuan and Cai (2010).

As a result of performing the eigendecomposition in FPCR, we only retain k_n vectors, which does not ensure continuity at the edges. In this situation, we then employ the penalized least squares fit to find a smoother version g_{SPCR} of g_{PCR} in $S_d^r(\triangle)$. We proceed as follows. We evaluate the discontinuous polynomial g_{PCR} at the domain points (Lai and Schumaker, 2007) that constitute the minimum set of points to fully characterize any bivariate polynomial of a fixed degree, by its function evaluations in the triangulation $\triangle$. Then we fit the continuous (or even smooth) spline g_{SPCR} that approximates best these function evaluations.

3 Numerical Results on the Prediction for Ozone Concentration Values

In this paper, we forecast the ground-level ozone concentration at one EPA station in suburban Atlanta (in Kennesaw, GA, latitude 34.0138°N, longitude 84.6049°N) using the random surfaces over the entire U.S. domain based on the measurements at various EPA stations from the previous days. The locations are scattered over a complicated domain, delimited by the United States frontiers, which has been scaled into $[0, 1] \times [0, 1]$, see Figure 1.

There are about 1000 EPA stations in the U.S. We are given the ozone concentrations at each locations for every hour over three months from July to September, 2005. The amount of the missing data is very small, so they affect very slightly our computations as the penalized least squares fit was used. Assume that the ozone concentration in a city, say Atlanta on one day at a particular time is a linear functional of the ozone concentration distribution over the surrounding region of the U.S. continent on the previous day. Also we may assume that the linear functional is continuous. Thus, we can use functional linear model for prediction.

With bivariate spline functions, we can easily carry out all the experiments to approximate the ozone distributed random surfaces and approximate the linear functional. Based on bivariate splines theory, the smaller the triangulation size, the better the approximation. In the following numerical experiments, we considered three different sizes of triangulations over the South East of the USA, depicted in Figure 2.

We first use BF, following (1), with penalty $\rho = 10^{-9}$. This level of penalty was considered reasonable after some preliminary analysis. We could have formally searched for an optimal penalty ρ but it was relatively clear that a few orders of magnitude of potential ρ values would work well, so we kept $\rho = 10^{-9}$. In that setting, $f(X)$ is the ozone concentrations at the station of interest at one particular hour of one day and X is the ozone concentration distribution function over the South East of the USA at the same hour, but on the previous day. We use a penalized least squares fit S_X of X with penalty γ to compute the empirical estimate $\widetilde{S_{\alpha,n}}$.

Let us explain our numerical procedures in detail. To forecast the ozone concentration on a particular day, we build hourly surfaces for all hours over all days in the sample, before the day at which we want to predict. We have two options: one in which 24 functions S_g are estimated using the data Y and X for each hour of the day (BF-1), and one in which one function S_g is estimated using all hours of the day (BF-24). In the second option, the sample size is 24 times larger, but the relationship between the surface of ozone at an hour and ozone at the station 24 hours ahead does not depend on the hour. This stationarity assumption weakens BF-24 but enables the computation of more robust predictions when the number of days in the sample is small, and these methods suffer from autocorrelation leakage for nearby hours. The same notations are adopted for FPCR-1 and FPCR-24. The methods BF-1 and FPCR-1 do no suffer from autocorrelation leakage as most studies show significant autocorrelation only over 2-3 hours for ground-level ozone, but use a much smaller sample size.

For illustration of both BF and FPCR, we display in Figure 3 the 24 hour ahead predictions over two different days, based on length varying from 10 to 22 days of learning. The methods give similar results, but FPCR seems to provide better forecasts than BF, with not as much of a need for a large sample size to provide good predictions. In the next two sections we quantify the predictive abilities of these two techniques.

Figure 4 and 5 respectively display estimates for the prediction functions g for BF-1 and FPCR-1 at four different times of the day (5AM, 11AM, 3PM and 6PM). The scales are different for BF and FPCR. Indeed, the empirical eigenvalues for FPCR are much larger here than for the matrix

$\widetilde{A}_n$ in BF, as we use polynomials without any smoothness condition across edges in a first step. The resulting eigenvectors that help us reconstitute the prediction functions g are therefore much smaller for BF.

3.1 BF predictions

Here we present numerical results using BF to predict at the Atlanta station. It is necessary to choose the value of γ. Figure 6 displays the averaged daily RMSEs of predictions over August 20-31, 2005 for various choices of γ, as well as averaged daily Mean Absolute Deviations (MAD) and averaged daily maximum error. It shows that RMSEs decrease rapidly with sample size and level off at around 20 observations. It also indicate that γ ought to not be chosen too large or too small, but its impact on the predictions is relatively minor for BF-24. For BF-1, it is necessary to choose a high value for γ, here $\gamma = 1$, as too much roughness in the hourly surfaces X can produce variability on the estimates of the functions g due to a much smaller sample size.

3.2 FPCR predictions

We show our numerical results using our FPCR approach to predict at the Atlanta station. In that setting, it is necessary to choose a number of eigenvalues/eigenvectors to carry out the predictions. Hence we have one additional tuning parameter in addition to the γ values. Figures 8 and 9 display RMSEs for one such choice of 2 eigenvalues according to sample size. It shows that for FPCR-24, RMSEs are already small for sample sizes of only 5 days, and decrease very slowly. It is a remarkable behaviour of this method to be able to learn with a very small number of days. For FPCR-1, which uses 24 times less data points but exactly at the hour of interest, the RMSEs are much larger but decrease steadily from these large values and reach lower values than FPCR-1 after around 20 days. Hence FPCR-1 is more tailored to the problem, but requires more days.

We considered 2, 4, 6 and 10 eigenvalues only, as forecasts were not improved by employing more than 10 eigenvalues. The tables of RMSEs ranked by size of the learning period, for 2, 4, and 6 eigenvalues, are not displayed here but show similar behavior. Table 1 summarizes RMSEs for these various choices for FPCR-24 and provides a comparison with BF-24 for a sample size of 30 days.

Such RMSE calculations can provide the basis for cross-validation over a subset of days prior to the prediction period. However, one must be careful to choose the adequate performance criterion (e.g. RMSE, MAD) to match cross-validation and the forecasting task (Gneiting, 2011). We used RMSE in the sequel but paid attention to potential variations across performance criteria, see Figure 6.

4 Comparison with thin-plate splines

In this section, we assess the predictive abilities of BF using either bivariate splines or thin-plate splines. Thin-plate splines also minimize $\sum_{k=1}^{N} |h(s_k) - X_i(s_k)|^2 + \gamma \text{En}(h)$ and lead to solution in a basis of thin-plate splines of the type:

$$h(x) = \sum_{i=1}^{n} \delta_i \eta(\|x - x_i\|) + \sum_{j=1}^{3} \alpha_j \phi_j(x)$$

where ϕ_j are linearly independent polynomials of degree 2, and $\eta(r)$ is proportional to $r^2 \log(r)$ (radial functions), see Wood (2003) and references therein. We use here the R package `fields` (Fields-Development-Team, 2006) to fit thin-plate splines. Figure 10 depicts the predictions for

10

September 10, with varying sample sizes, obtained using the BF with thin-plate splines (with penalty *gamma* chosen by cross-validation) instead of bivariate splines. A lack of coherence in the predictions across hours can be seen when small samples are used to predict. For instance, forecasts at 18:00, 19:00 and 20:00 dramatically differ for sample sizes of 10 and 14 days, and dampen for larger sample sizes. This phenomenon occurs systematically over the validation period. The explanation is that the lack of spatial fit, compared to bivariate splines, may deteriorate the inference when only limited information is available. We can quantify this variability through the use of jackknife (Efron, 1979). For a day, we carry out predictions using 10 out of the 11 days previous to the day before predictions are made, removing one day at a time out of the sample, and we produce 10 predictions for the BF with either bivariate splines or thin-plate splines. As shown in Figure 11, the resulting variability in the forecasts is much larger using thin-plate splines compared to using bivariate splines.

Table 2 shows that the BF-1 with bivariate splines (BF-BS-1) surpasses the BF-1 with thin-plate splines (BF-TPS-1) over the period September 5-14 (not overlapping the period September 1-4 considered in the jackknife illustration above and showing similar behavior). It is so by a large margin for small sample sizes and by a small margin for larger sample sizes. The BF with thin-plate splines needs at least 16 days of learning to give reasonable predictions, due to the aforementioned high sensitivity to the given sample. The decrease in RMSE is slow for 10-15 days and then abrupt from 15 to 16. This pattern is reflected as well over one day in Figure 10 where we can observe a sudden improvement in predictive power for the morning (mean level) as well as the afternoon (large reduction in variability); the possible explanation for such sensitivity may be the non-linear nature of the thin-plate splines fitting compared to bivariate splines fitting. Table 2 also shows the advantage of using 24 hours when computing the estimate of g: one does not need a large number of days to be able to produce reasonable predictions. However, the drawback of this feature is that when sample size increases, the predictions are made using too much averaging, compared to the tailored 1-hour based version, as seen clearly when comparing the variation of RMSEs with sample size for BF-BS-24 and BF-BS-1. To compensate for larger instability when using smaller sample size in BF-BS-1 and FPCR-1, we increased the penalty up to $\gamma = 1$ to obtain good predictions.

Thin-plate splines are globally supported, so large scale influences can occur in the fit with no respect of boundaries and sharp variations. Bivariate splines are locally supported, and can fit in any geometry of polygonal domain of interest. As a result, bivariate splines are more likely to better represent local variations. Furthermore, the computational cost of using radial functions of the kind $r^2 \log(r)$ is large due to numerical approximations in the integrals used to derive scalar products, whereas bivariate splines are merely polynomials for which immediate and exact results for integrals can be obtained. We noticed differences of several orders of magnitude in our particular computer set-up between bivariate splines and thin-plate splines. To overcome the difficulty of numerical approximations, Reiss and Ogden (2010) used another basis of radial functions: radial cubic B-splines over compact domains for a neuroimaging application. However, these compactly supported splines do not possess optimal approximation properties, so their use may be restricted to sample locations that are numerous and spatially covering the domain of interest, as in neuroimaging.

5   Comparison with the VAR model

In Table 2, we compare the performance of our functional approach with a standard multivariate time series approach: the vector autoregression model (VAR). We use the the R package `vars` (Pfaff, 2008) with the default settings. We first select each vector of EPA stations in the South East at a selected hour. There are 210 locations in our data set, and fitting the VAR requires at

least 210 observations, so we cannot use a model for each hour of the day as for BF-1 and FPCR-1. Therefore we need to make use of all the previous hours in the 10-17 day samples to have enough information to fit the VAR. To make a prediction using 10 days of learning for September 10, we are using a sample size of 240 hours, which is the same as in the BF/FPCR-24 models.

The next challenge in employing the VAR model is to impute a few missing data. To do so, we use either the spatial average over all available stations at the time when the observation is missing or the average value over all available observations at the same hour and at the same location where the data is missing. We denote these methods VAR-space and VAR-time in Table 2. Note that bivariate spline methods do not suffer from these imputation problem as they naturally include an excellent interpolation step. We could have employed this step in the VAR-space to impute these few missing values but this was unlikely to significantly change the VAR predictions and we sticked to the standard VAR approach. Finally, because observations are rounded to the next integer by the EPA, many of the columns in the observations are collinear, so we add negligible normally distributed random numbers with mean zero and standard deviation 0.01 to each observation to make the system solvable (correcting any values that becomes negative). The VAR models have huge errors for the 10 day sample size because the system is nearly collinear. However, we do see a decrease in errors as the sample size increases. Nevertheless, both VAR approaches yield RMSEs around 40-100% higher than the BF and FPCR for the maximum sample size of 17 days considered here.

6   Effect of triangulation

Unsurprisingly, preliminary analysis showed that large scale triangulations (such as the entire continental USA) do not produce good forecasts as these spatial scales are producing spurious effects. Indeed, the chemistry and transport occur only at a regional scale on the 24 hour-ahead horizon. Hence, we only considered triangulations (see Figure 2) over the South-East USA to carry out statistical inference. We were not able to draw a conclusion in terms of which one is the best as none of these three triangulations is consistently better than the others across the days of the validation period. Obviously, there is an infinite number of triangulations for this region. Finding an optimal triangulation for our purpose is very interesting and could lead to even better predictive abilities. It seems to be better to use a triangulation dependent on the geography as one wants to borrow strength spatially in a meaningful manner. For example, the ridges of mountains should be edges of the triangulation; cities should be taken into account as predictions, for instance by putting them at the vertices of the triangulation. This investigation is beyond the scope of this paper.

7   Conclusions and Future Research

The BF method works very well for Atlanta. The predictions are consistent for various learning periods and the predictions are close to the exact measurements. Using the first two eigenvalues or more, FPCR also works very well. It is hard to say which one is better in general. However, on average, FPCR seems to slightly outperform BF in our example as it requires a smaller sample size to be efficient.

Overall, we recommend BF since it is simpler than FPCR which requires some expert knowledge about how many eigenvalues are needed for the best prediction. In our examples, we observed the decay of the eigenvalues, and often could not find an abrupt fall, also called a knee, if the values of all eigenvalues are plotted a normal scale. When plotted on a logarithmic scale, each of eigenvalues (in the first 10 of them) has a knee. It is hard to decide which one is the right knee. Determination

of how many eigenvalues for the best prediction is not an easy task and requires further study. Yuan and Cai (2010) showed not only that the FPCR requires additional assumptions to be valid, but that on well-designed simulations, predictions given by BF, with penalty ρ chosen with cross-validation, compares favorably to the FPCR from Hall and Horowitz (2007). However, if one has information about the principal components, or if the situation is made easy for their estimation, it may be possible for FPCR to outperform BF in these cases. A reasonably accurate estimation of principal components might explain why this occurred in our example over the prediction period September 5-14. Finally, Ferraty et al. (2011) considered pre-smoothing for FPCR, that is the perturbation of the normal equation $\Gamma g = \Delta$ at the beginning of the analysis. It seems to significantly improve the FPCR when the sample size is small or the noise is large. One could expect our results to be enhanced accordingly with the use of presmoothing.

It is also interesting to study the ozone value prediction at other cities, e.g., Boston, Cincinnati and others, see Ettinger (2009) for more details. The numerical results again show that BF is easy to use and performs well. We see that the quality of predictions reaches a plateau after a learning period of around 15 days. Although in theory we can predict better if we use a longer learning period, the numerical results vary based on our experiments. We also conducted experiments for predictions using bivariate splines of degree d bigger than 5: we employed $d = 6, 7, 8$ and 9 (not reported here). The numerical results are broadly similar to the predictions using bivariate splines of degree $d = 5$.

The inclusion of co-variates (e.g. meteorological or even chemistry-transport model predictions) can improve such purely statistical ozone predictions (Damon and Guillas, 2002; Guillas et al., 2008). We showed here the abilities of our method with no covariates, with potential gains from the use of this additional information. However, the treatment of covariates (themselves spatially distributed) in this spatial functional data context is challenging. One possibility is first to merely add a non-spatially varying average effect as a parametric component. Such semiparametric functional models are now well understood (Aneiros and Vieu, 2006). A better idea would be to fully integrate the functional covariates. Modelling and computational issues arise, and this is currently under investigation.

Crainiceanu and Goldsmith (2010) recently developed Bayesian approaches for functional data, including a specific treatment of FPCR. Uncertainties are naturally computed as a result. Such Bayesian methods would most probably improve the quantification of uncertainties.

When the overall geographical coverage of triangulation is reasonably small (the coarseness is another issue), the predictions are better for both BF and FPCR for our particular example. Indeed, one should restrict itself to regions for which the spatial variability, through chemistry and transport, corresponds to the time scale of the prediction. As a result, larger regions, possibly the entire continental US, may be appropriate for predictions at longer horizons when larger scale transport is occuring. One can foresee adaptive triangulations that would suit particular conditions. The major conclusion from our results above is that bivariate splines outperform thin-plate splines in this context both in precision and computing time, and would do even more when the domain shows further constraints such as mountains and sea-shore that have a large impact on chemistry and transport due to their local features. Such qualities are attractive in various settings where taking boundaries into account would help improve interpolation over thin-plate splines (Newlands et al., 2011). We believe that our method has the potential to be implemented for operational pollution forecasting as a result.

References

Aneiros, G. and P. Vieu (2006). Semi-functional partial linear regression. *Statist. Probab. Lett. 76*(11), 1102–1110.

Aneiros-Perez, G., H. Cardot, G. Estevez-Perez, and P. Vieu (2004). Maximum ozone concentration forecasting by functional non-parametric approaches. *Environmetrics 15*, 675–685.

Awanou, G., M.-J. Lai, and P. Wenston (2006). The multivariate spline method for numerical solution of partial differential equations and scattered data interpolation. In *Wavelets and Splines: Athens 2005, G. Chen and M.-J. Lai (eds)*, pp. 24–74. Nashboro Press.

Bell, M., A. McDermott, S. Zeger, J. Samet, and F. Dominici (2004). Ozone and Short-term Mortality in 95 US Urban Communities, 1987-2000. *J. Am. Med. Assoc. 292*(19), 2372–2378.

Cai, T. and P. Hall (2006). Prediction in functional linear regression. *Ann. Statist. 34*(5), 2159–2179.

Cardot, H., F. Ferraty, and P. Sarda (1999). Functional linear model. *Stat. Probab. Lett. 45*, 11–22.

Cardot, H., F. Ferraty, and P. Sarda (2003). Spline estimators for the functional linear model. *Stat. Sin. 13*, 571–591.

Crainiceanu, C. and A. Goldsmith (2010). Bayesian Functional Data Analysis Using WinBUGS. *J. Stat. Softw. 32*(11), 1–33.

Crambes, C., A. Kneip, and P. Sarda (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist. 37*(1), 35–72.

Cressie, N. and C. K. Wikle (2011). *Statistics for Spatio-Temporal Data*. Wiley.

Damon, J. and S. Guillas (2002). The inclusion of exogenous variables in functional autoregressive ozone forecasting. *Environmetrics 13*, 759–774.

Dou, Y., N. Le, and J. Zidek (2010). Modeling hourly ozone concentration fields. *Ann. Appl. Stat. 4*(3), 1183–1213.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist. 7*(1), 1–26.

Ettinger, B. (2009). *Bivariate Splines for Ozone Density Predictions*. Ph. D. thesis, Univ. of Georgia.

Ferraty, F., W. González-Manteiga, A. Martinez-Calvo, and P. Vieu (2011). Presmoothing in functional linear regression. *Statist. Sinica*. to appear.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. London: Springer-Verlag.

Fields-Development-Team (2006). fields: Tools for spatial data. Technical report, National Center for Atmospheric Research, Boulder, CO, USA.

Gneiting, T. (2011). Making and evaluating point forecasts. *J. Am. Stat. Assoc.*. to appear, avalaible at `arXiv:0912.0902v2`.

Guillas, S., J. Bao, Y. Choi, and Y. Wang (2008). Statistical correction and downscaling of chemical transport model ozone forecasts over Atlanta. *Atmospheric Environment 42*, 1338–1348.

Guillas, S. and M. Lai (2010). Bivariate splines for spatial functional regression models. *J. Nonparametric Stat. 22*(4), 477–497.

Hall, P. and J. Horowitz (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist. 35*(1), 70–91.

Lai, M.-J. (2008). Multivariate splines for data fitting and approximation. In *Approximation Theory XII: San Antonio 2007, M. Neamtu and L. L. Schumaker (eds.)*, pp. 210–228. Nashboro Press.

Lai, M.-J. and L. Schumaker (2007). *Spline Functions over Triangulations*. Cambridge University Press.

Neidell, M. (2010). Air quality warnings and outdoor activities: evidence from Southern California using a regression discontinuity design. *J. Epidemiol Community Health 64*(10), 921–926.

Newlands, N. K., A. Davidson, A. Howard, and H. Hill (2011). Validation and inter-comparison of three methodologies for interpolating daily precipitation and temperature across canada. *Environmetrics 22*(2), 205–223.

Paciorek, C. J., J. D. Yanosky, R. C. Puett, F. Laden, and H. H. Suh (2009). Practical large-scale spatio-temporal modeling of particulate matter concentrations. *Ann. Appl. Stat. 3*(1), 370–397.

Pfaff, B. (2008). Var, svar and svec models: Implementation within R package vars. *Journal of Statistical Software 27*(4).

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis*. Springer-Verlag.

Reiss, P. and R. Ogden (2010). Functional Generalized Linear Models with Images as Predictors. *Biometrics 66*(1), 61–69.

Wood, S. N. (2003). Thin plate regression splines. *J. R. Stat. Soc. Ser. B Stat. Methodol. 65*(1), 95–114.

Yuan, C. and T. Cai (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Stat. 38*(6), 3412–3444.

Tables and Figures

	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
2	11.61	11.42	11.34	11.35	11.25	11.40	12.40	14.92
4	11.60	11.21	11.21	11.35	11.46	11.68	12.07	14.28
6	12.03	11.23	11.22	11.51	11.29	11.58	12.10	12.97
10	12.77	11.23	11.36	11.46	11.24	11.68	12.24	12.23
BF	11.33	11.20	11.03	10.87	11.00	10.86	10.84	11.57

Table 1: Average 30 Day RMSE over August 20-31 by γ values. FPCR-24 for 2, 4, 6 and 10 eigenvalues, BF-24 for $\rho = 10^{-9}$.

	BF-TPS-1	**BF-BS-24**	**BF-BS-1**	**FPCR-24**	**FPCR-1**	**VAR-space**	**VAR-time**
10	19.75	13.68	14.58	9.82	10.08	3.07×10^8	4.87×10^9
11	19.46	13.58	13.61	9.91	9.98	50.78	111.73
12	19.21	13.22	12.70	9.91	9.94	33.18	37.54
13	18.76	13.03	12.12	9.90	9.99	27.70	28.36
14	18.32	12.94	11.52	9.89	9.87	23.41	40.78
15	18.36	13.15	11.30	9.91	9.74	18.80	23.16
16	12.48	12.95	11.05	9.92	9.82	20.40	20.51
17	12.08	12.55	10.98	9.85	9.80	19.79	17.77

Table 2: Average RMSEs of BF with thin-plate splines (BF-TPS), BF-24 with bivariate splines ($\gamma = 10^{-4}$), BF-1 with bivariate splines ($\gamma = 1$), FPCR-24 with bivariate splines ($\gamma = 10^{-4}$), FPCR-1 with bivariate splines ($\gamma = 1$), VAR-space, VAR-time, over Sept 5-14 (10 day-long validation period) with sample sizes from 10 to 17 days.
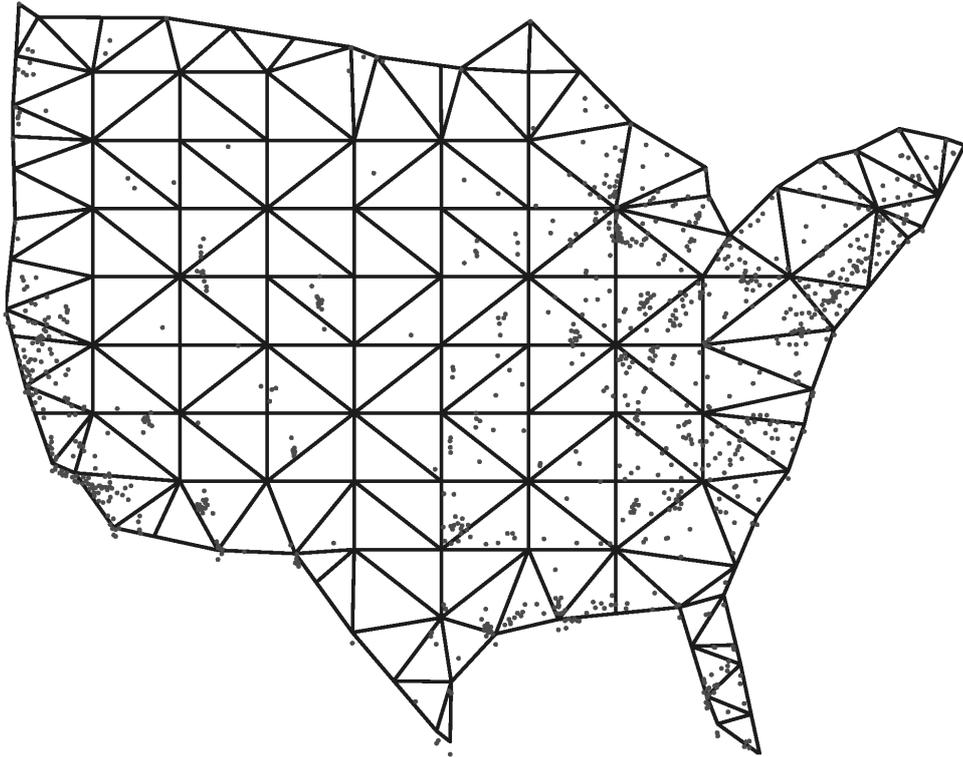
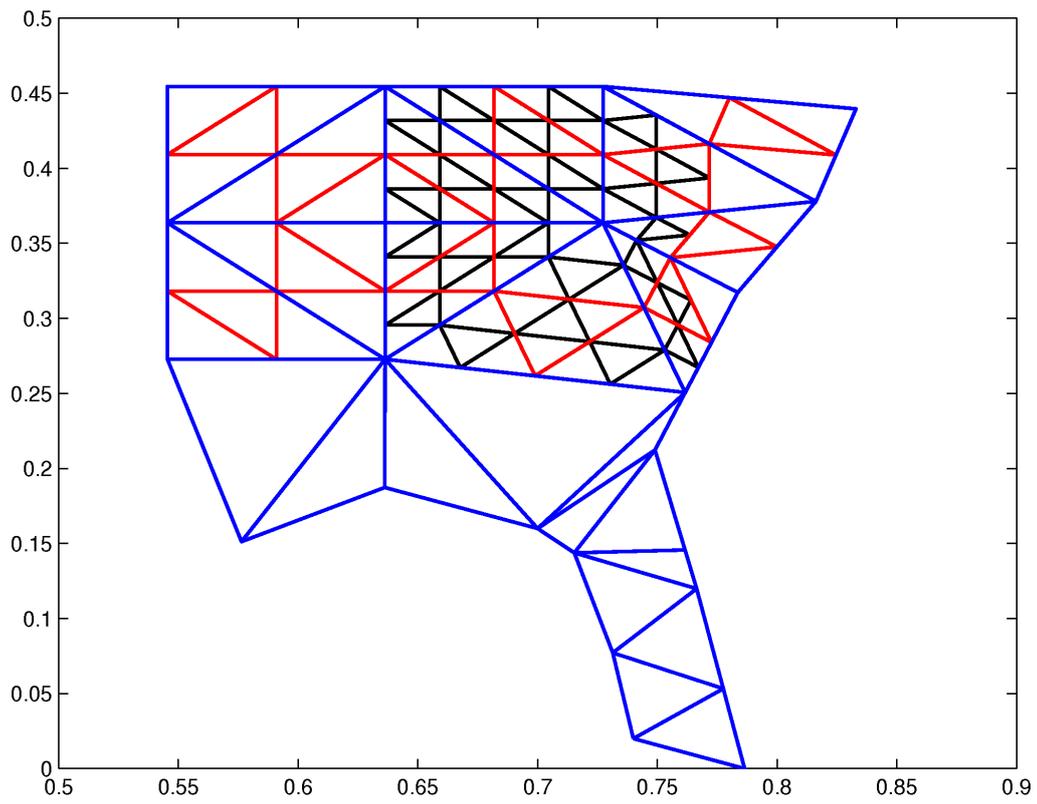Figure 1: Locations of EPA stations and an example of a triangulation

Figure 2: Triangulations T_1 (blue), T_2 (red), T_3 (black) of the South East of the USA.
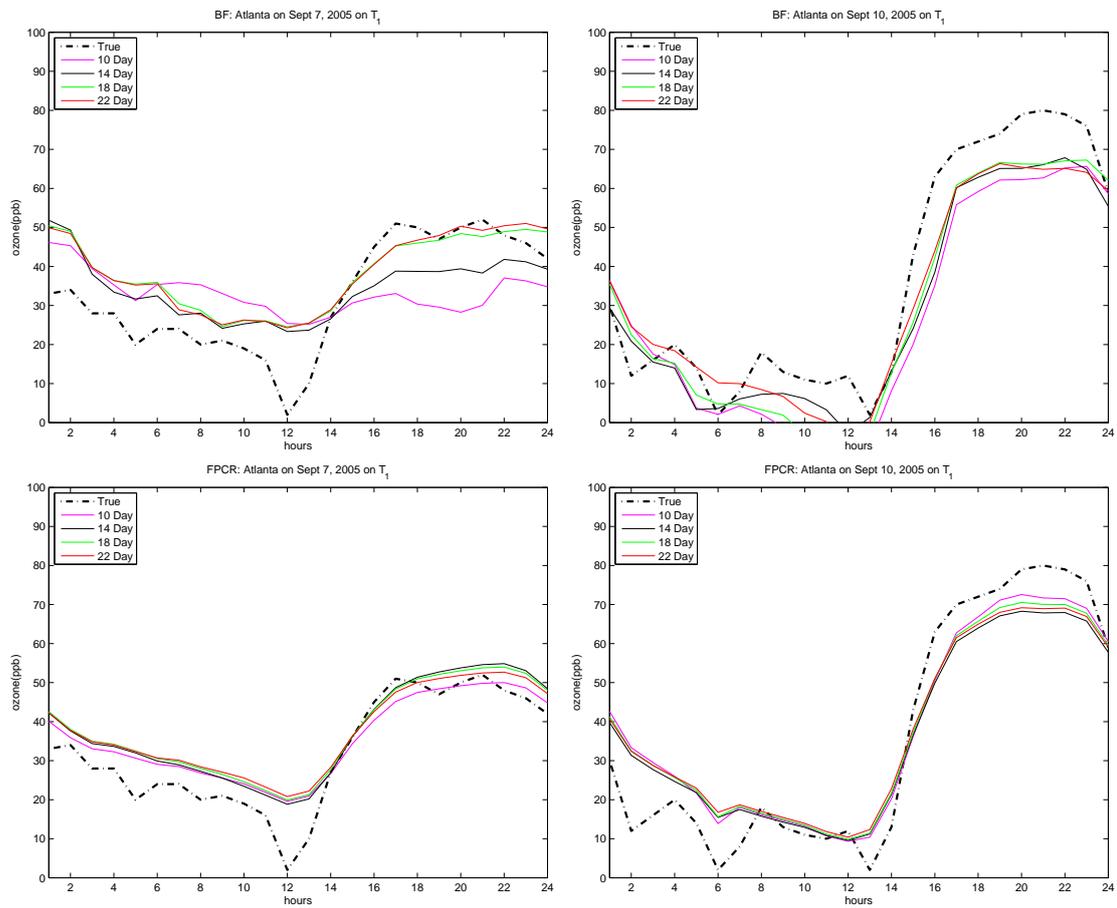
Figure 3: Predictions using BF and FPCR with bivariate splines, over Sept. 7 and Sept. 10, with varying sample sizes for the learning periods prior to predictions, penalty $\gamma = 10^{-4}$ for both methods and two eigenvalues for FPCR.
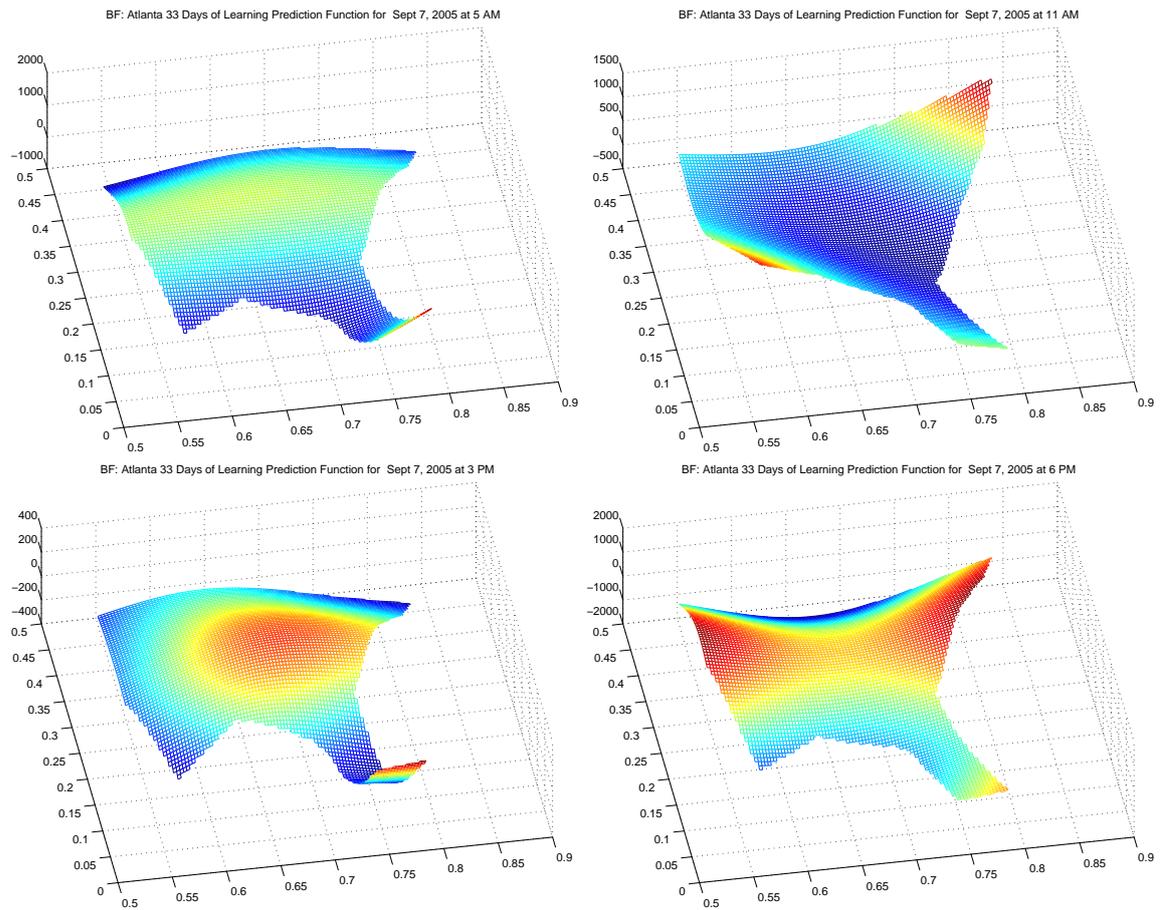
Figure 4: Estimated prediction functions g at different times for BF-1, penalty $\gamma = 10^{-4}$
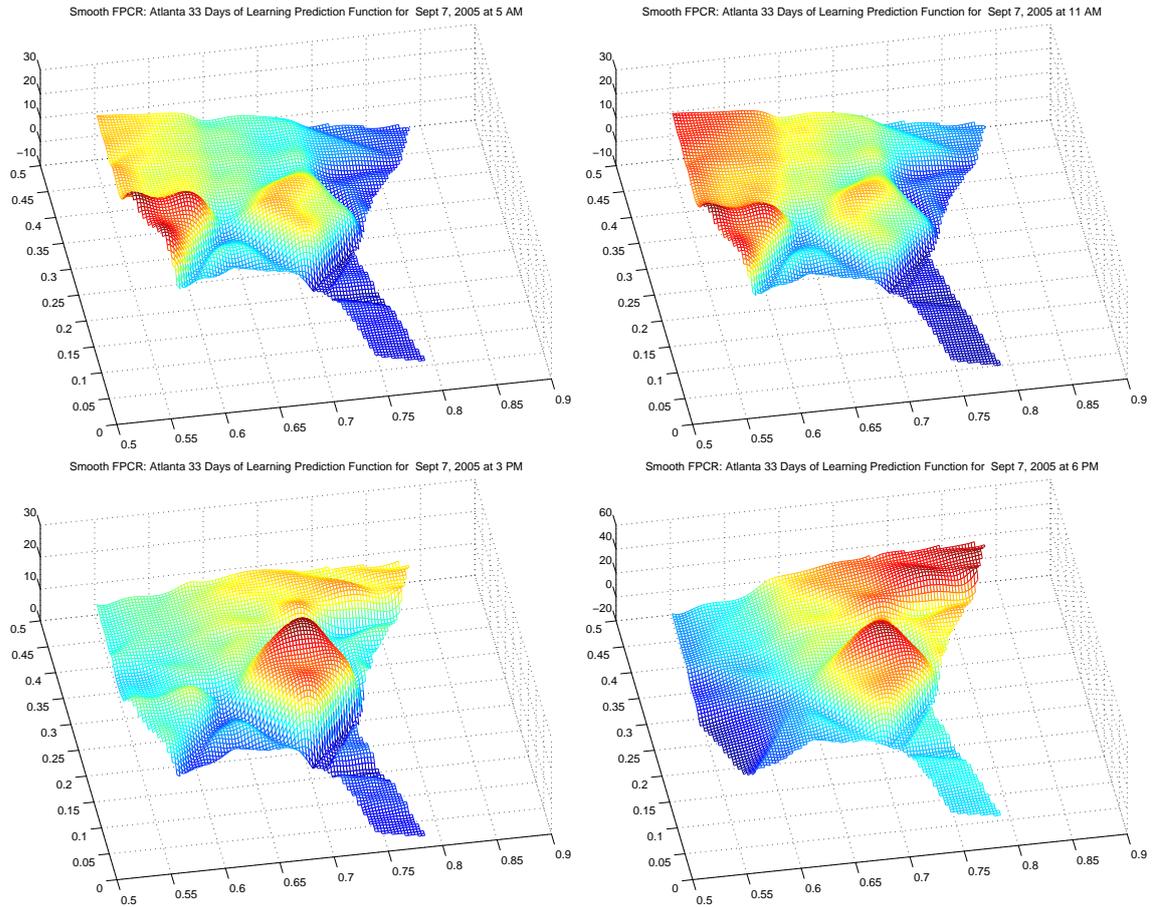
Figure 5: Estimated prediction functions g at different times for FPCR-1, penalty $\gamma = 10^{-4}$ and two eigenvalues.

22
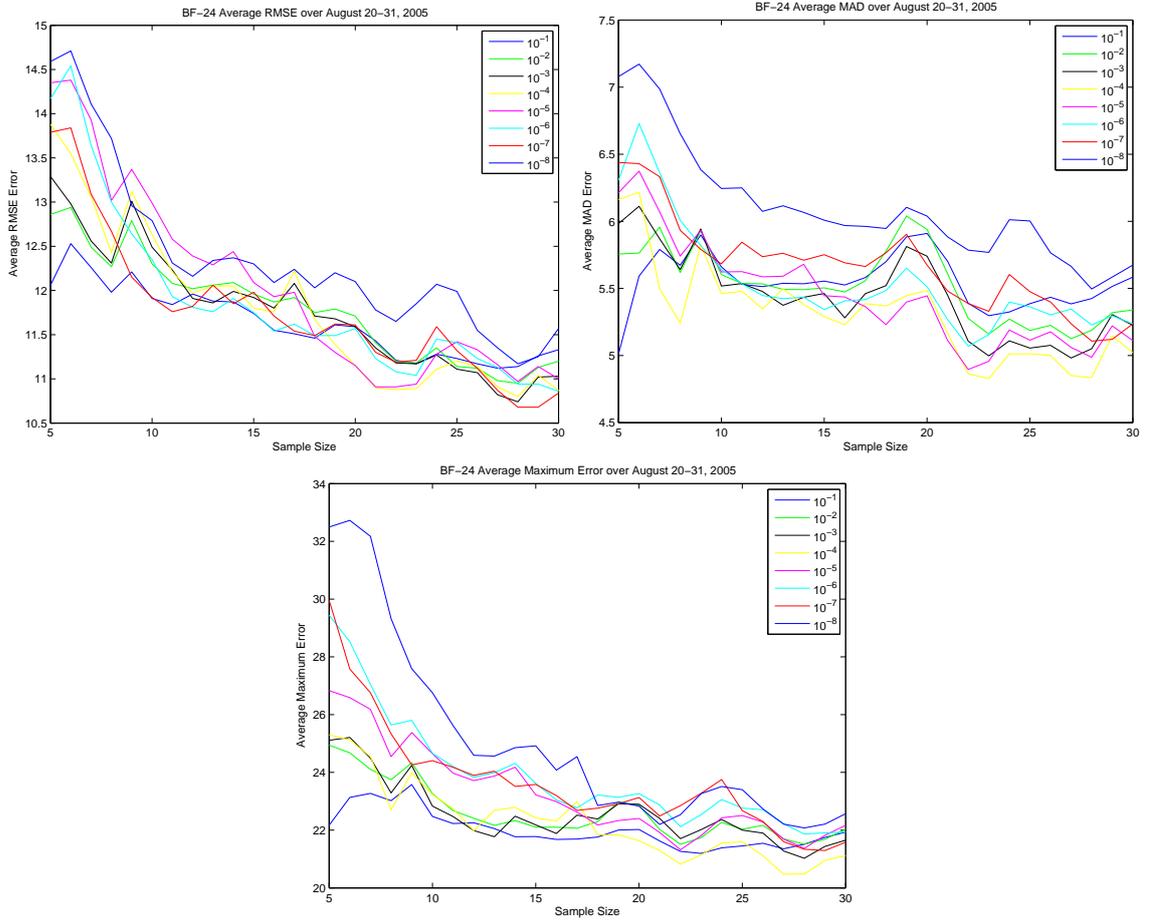
Figure 6: BF-24: Averaged daily RMSE, MAD and maximum error over August 20-31, 2005. Sample Size varies from 5 days of learning to 30 days, and penalty for bivariate splines fitting γ varies in the range 10^{-1} to 10^{-8}.
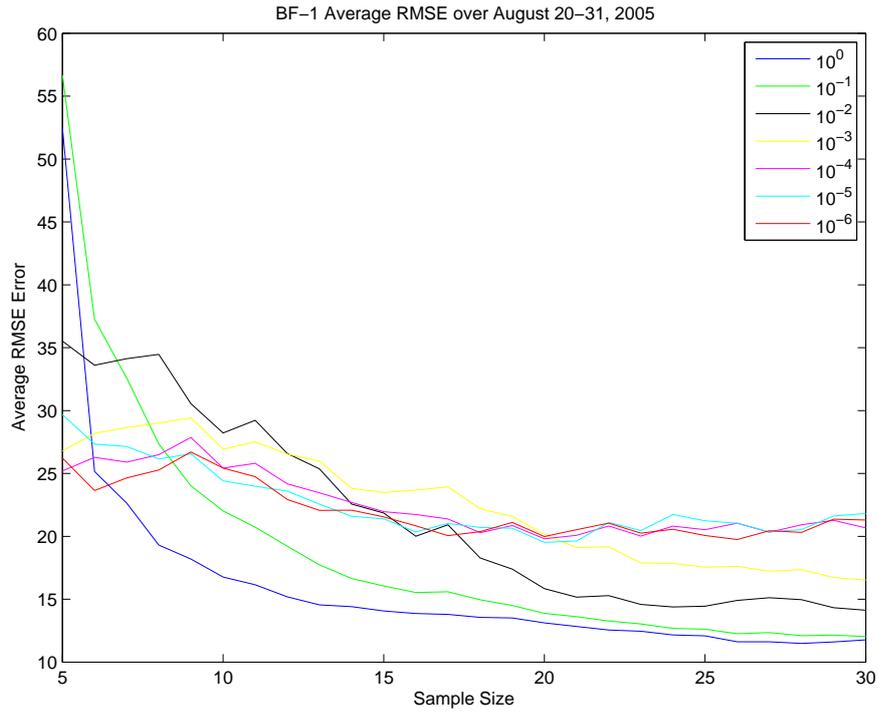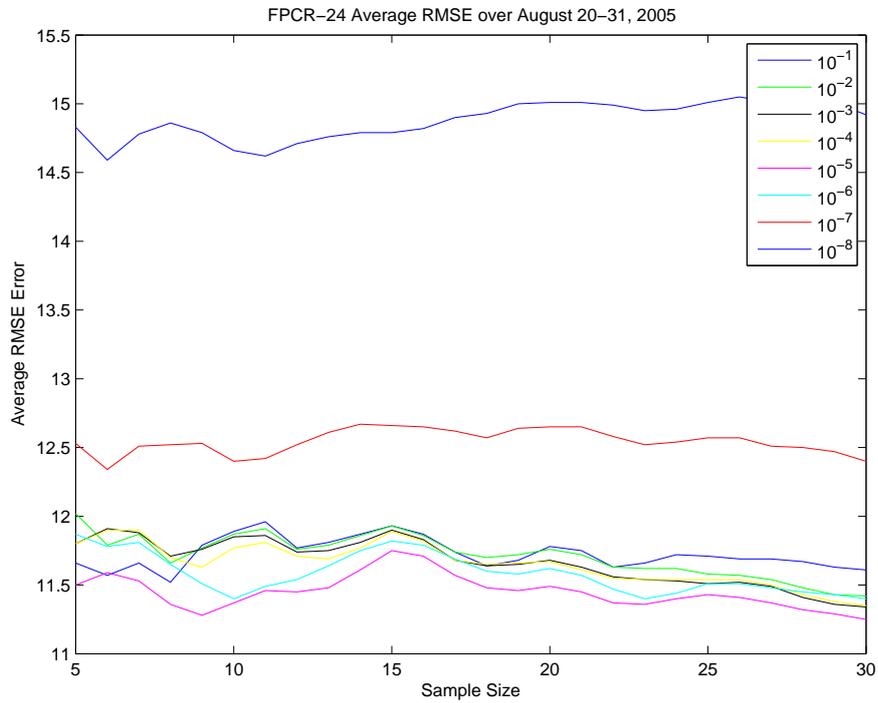
Figure 7: BF-1 Average RMSE over August 20-31, 2005. Sample Size varies from 5 days of learning to 30 days, and penalty for bivariate splines fitting γ varies in the range 10^0 to 10^{-6}.

Figure 8: FPCR-24 average RMSE over August 20-31 and 2 Eigenvalues. Sample Size by γ values, with number of days for the training period between 5 and 30.
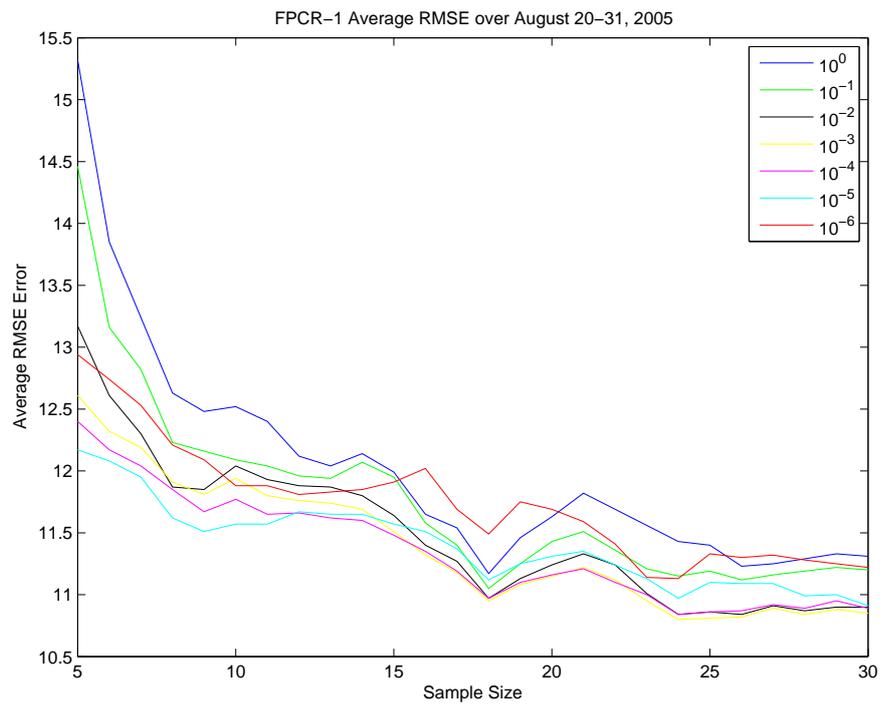
Figure 9: FPCR-1 average RMSE over August 20-31 and 2 Eigenvalues. Sample Size by γ values, with number of days for the training period between 5 and 30.
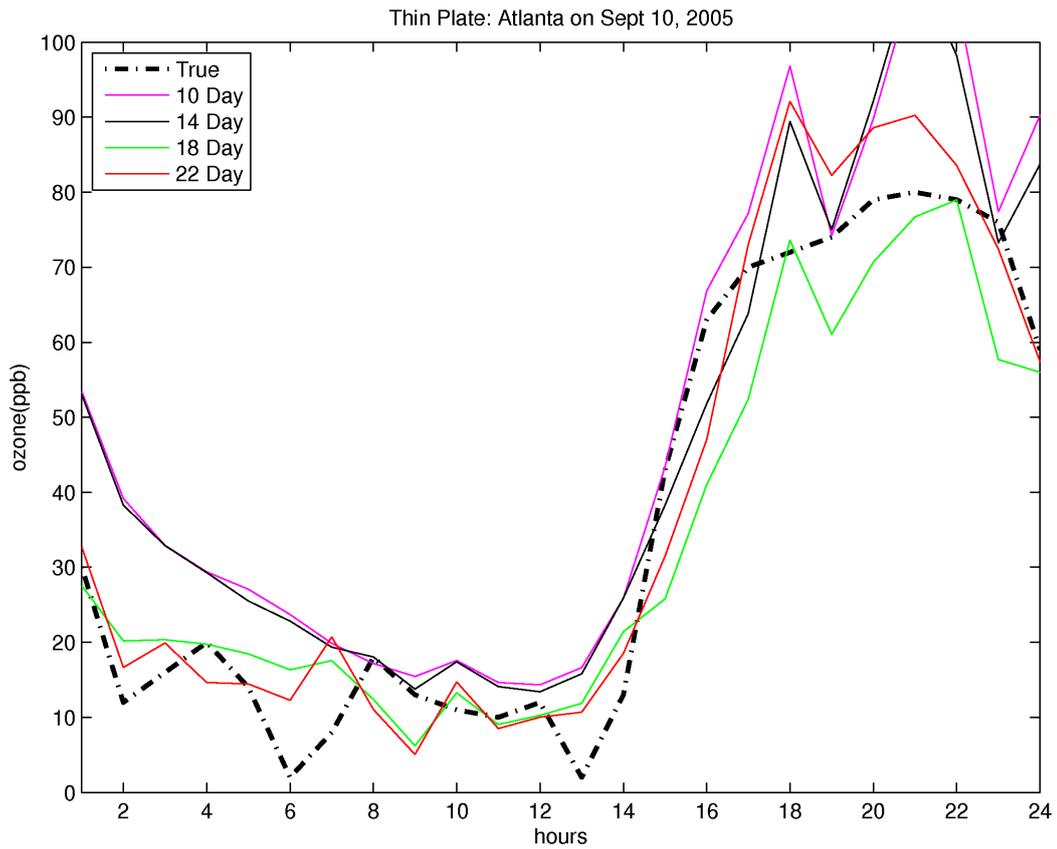
Figure 10: Prediction using BF with thin-plate spline (BF-TPS), over Sept 10, with varying sample sizes for the learning periods prior to predictions.
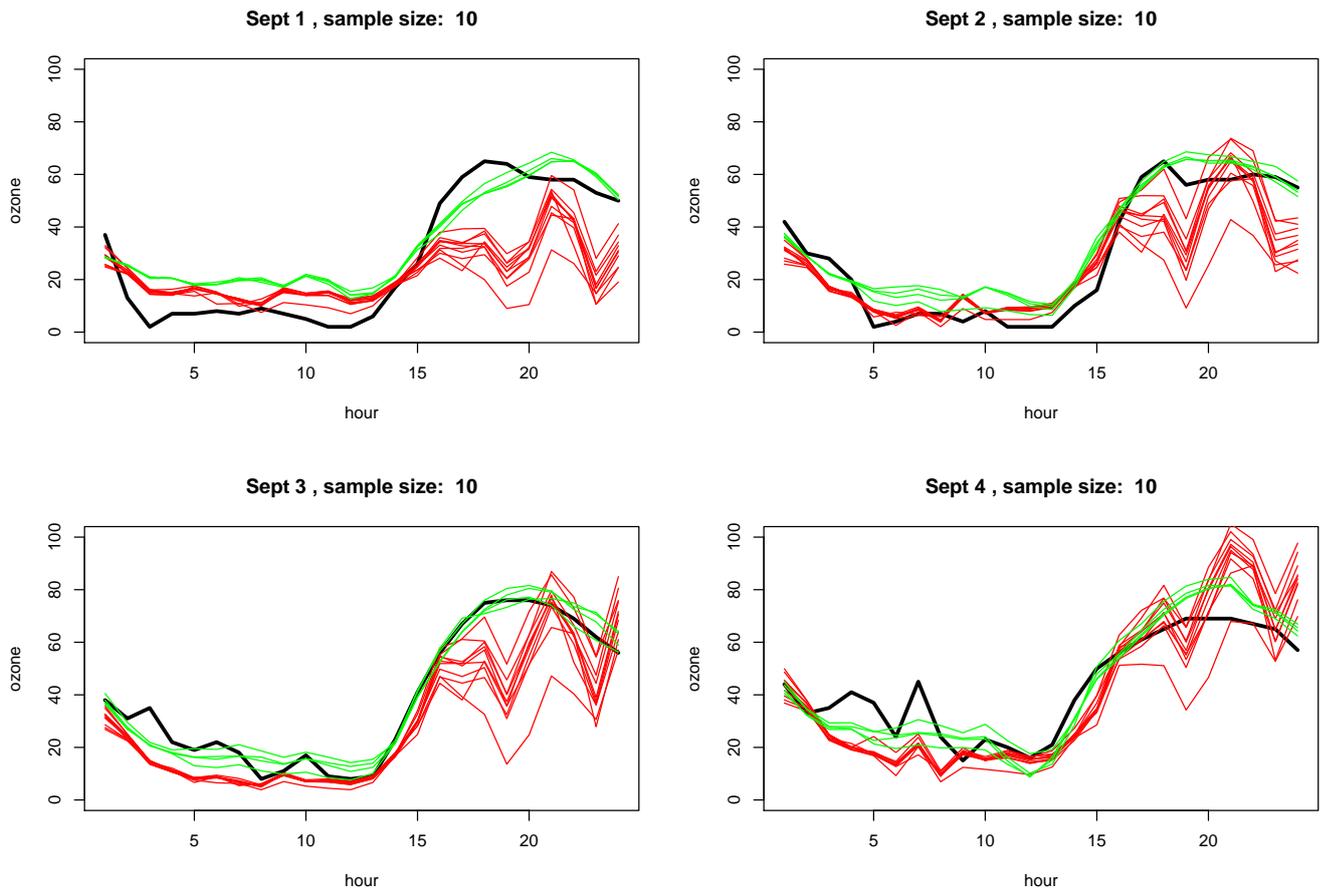
Figure 11: Brute force jackknife predictions (sample size 10, after leave one out of 11), with bivariate splines (green), thin-plate splines (red) over September 1 to 4.