Semiparametric Geographically Weighted Response Curves with Application to Site-Specific Agriculture

Scott HOLAN, Suojin WANG, Ali ARAB,
E. John SADLER, and Kenneth STONE

Lack of basic knowledge about spatial and treatment varying crop response to irrigation hinders irrigation management and economic analysis for site-specific agriculture. One model that has been postulated for relating crop-specific economic quantities to irrigation is a quadratic response curve of yield as a function of irrigation. Although this model has far reaching economic interpretations it does not account for spatial variation or possible nitrogen–irrigation interactions. To this end we propose a spatially treatment varying coefficient model that alleviates these limitations while providing measures of uncertainty for the estimated coefficient surfaces as well as other derived quantities of interest. The modeling framework we propose is of independent interest and can be used in many different applications. Finally, an example involving site-specific agricultural data from the U.S. Department of Agriculture–Agricultural Research Service demonstrates the applicability of this methodology.

Key Words: Additive models; Bivariate smoothing; Penalized splines; Semiparametric regression; Varying coefficient models.

1. INTRODUCTION

Sadler, Camp, Evans, and Millen (2002) conducted an experiment with the goal of measuring the mean response of corn to irrigation amounts on 12 soil map units. Additionally, this research compares the variation in the response within and among soil map units. During the course of this research it was determined that corn yield can be described as a quadratic function of irrigation. Such a production function (a.k.a. response curve) is common in irrigation research and provides the practitioner with the essential information needed to make appropriate water applications and achieve optimal crop response. In order

Scott Holan is Assistant Professor, Department of Statistics, University of Missouri–Columbia, 146 Middlebush Hall, Columbia, MO 65211-6100 (E-mail: *holans@missouri.edu*). Suojin Wang is Professor, Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143. Ali Arab is Assistant Professor, Department of Mathematics, Georgetown University, 305 St. Mary's Hall, Washington, DC 20057-1233. E. John Sadler is Research Leader, USDA–ARS Cropping Systems and Water Quality Research Unit, RM 269 Ag. Eng. Bldg, University of Missouri–Columbia, MO 65211-6100. Kenneth Stone is Agricultural Engineer, USDA-ARS Coastal Plains Soil, Water and Plant Research Center, 2611 West Lucas St., Florence, SC 29501.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 13, Number 4, Pages 424–439
DOI: 10.1198/108571108X378308

to fully take advantage of the quadratic relationship between crop yield and irrigation, a model that makes use of the inherent spatial dependence is required.

Varying coefficient models are concerned with a class of regression or generalized regression models whose coefficients are allowed to vary as smooth functions of other variables and were first developed by Hastie and Tibsharani (1993). Since the models' first inception, there have been many research efforts leading to numerous methodological advances. Although varying coefficient models have become an active area of research, by comparison, there have been relatively few efforts aimed at models having spatially varying coefficients.

One article that addresses the topic of spatially varying coefficient models is Assunçao, Gamerman, and Assunçao (1999). In this research the authors introduced a Bayesian model with space-varying parameters in order to study microregion factor productivity in Brazilian agriculture. Another contribution to this area can be attributed to Gamerman, Moreira, and Rue (2002) who developed a modeling approach for space-varying regression. However, the model they proposed makes the assumption that a given coefficient is constant over specified areal units, rendering the model somewhat restrictive. In order to circumvent this inflexibility one could introduce parametric models for each coefficient. This type of methodology was fully described by Fotheringham, Brunsdon, and Charlton (2002) where the authors provided a comprehensive treatment of the subject of geographically weighted regression.

Luo and Wahba (1998) used a spline surface model over a two-dimensional space in the context of meteorological data. The modeling technique they developed provides greater flexibility but requires the user to choose the spline basis function along with the number and locations of the knot points. In the context of locally stationary spatial modeling Agarawal, Gelfand, Sirmans, and Thibadeau (2003) introduced local regression models for assessing factors that control housing prices in school (and subschool) district levels.

In the general research area of spatially varying coefficient models, several Bayesian approaches have also been proposed. One early research effort that used spatially varying coefficients is Wikle, Berliner, and Cressie (1998). This research proposes a hierarchical model for monthly surface temperatures in the Midwest. Here parameters controlling the mean, annual cycle, and vector autoregressive (VAR) dynamics are allowed to be spatially varying. Additionally, Assunçao, Potter, and Cavenaghi (2002) provided a Bayesian method for generalized linear models with coefficients allowing for spatial dependence. Further, Gelfand, Kim, Sirmans, and Banerjee (2003) built regression models to explain a response variable over a region of interest under the assumption of spatial dependence. More recently, Wikle and Anderson (2003) considered a spatio-temporal, zero-inflated Poisson (ZIP) model. The model they proposed allows coefficients for climate to vary with space. In addition, since the dimensionality was quite high, the coefficients were modeled as an expansion in terms of a lower-dimensional set of spatial basis functions. Penalized splines and generalized additive models, for analyzing spatial temporal data, were proposed by Fahrmeir, Kneib, and Lang (2004), although this work also adopts a Bayesian approach. Finally, a recent implementation of Bayesian spatially varying growth curve models with application to weed growth was proposed by Banerjee and Johnson (2006).

Although our application of interest can be formulated using a Bayesian approach, that is not the main focus of our analysis. However, we feel the effectiveness and suitability of such methods warrant their inclusion. Moreover, since the link between mixed models and Bayesian methods are straightforward and may be of interest to a portion of the intended audience we provide a brief description of this methodology. Of course, in addition to producing flexible, geographically weighted response curves, one of the explicit goals of the analysis we conduct is simplicity of exposition. Specifically, we wish to provide methods that are easily accessible to practitioners (agronomists) having a background only in regression. Therefore, a Bayesian approach is only briefly described. For a more thorough treatment of Bayesian semiparametric regression see, for example, Crainiceanu, Ruppert, and Wand (2005); Zhao, Staudenmayer, Coull, and Wand (2006); Giminez, Barbraud, Crainiceanu, Jenouvrier, and Morgan (2006); and the references therein.

The research at hand is motivated by the need to model spatially dependent response curves, specifically for application to site-specific agriculture. Additionally, when constructing response curves, it is often the case that the response curve varies according to the level of an additional treatment, forcing the practitioner to construct several different curves during the course of the analysis. Therefore, in addition to allowing the response curve to vary spatially, it is also of general interest to implicitly allow the response curve to vary according to the level of an additional factor (treatment).

In this article we propose semiparametric weighted response curves that vary over both space and treatment level. That is, the coefficients in our response curve allow for spatial dependence as well as a possible treatment A by treatment B interaction. Specifically, we model the coefficient surfaces using penalized spline regression, a nonparametric smoothing method that has gained recent popularity since the seminal work of Eilers and Marx (1996). Furthermore, the semiparametric model we propose exploits the equivalence between penalized splines and mixed models. This framework benefits from its ease of implementation while providing tremendous flexibility; see Ruppert, Wand, and Carroll (2003, chap. 1) for a comprehensive discussion. Finally, we construct approximate (point-wise) confidence intervals for the coefficient surfaces conditional on a given treatment level as well as for several researcher-defined (derived) quantities of interest.

This article is organized as follows. Section 2 describes the motivating application, analysis of site-specific agricultural data from the U.S. Department of Agriculture–Agricultural Research Service. Section 3 provides details surrounding the proposed methodology while Section 4 describes *semiparametric geographically weighted response curves* both in general and as it applies to our motivating example. A Bayesian implementation is presented in Section 5. Finally, concluding remarks are provided in Section 6.

2. SITE-SPECIFIC AGRICULTURE EXPERIMENT

The experiment we consider was conducted during the 1999–2001 corn growing seasons at the site-specific center-pivot irrigation facility in Florence, South Carolina. Although the data were analyzed over all three years, each year was analyzed separately. To this end, as an illustration of the proposed methodology, we present a contemporaneous analysis from 1999.

The site of the experiment was mapped on a 1:1200 scale by USDA-SCS staff in 1984

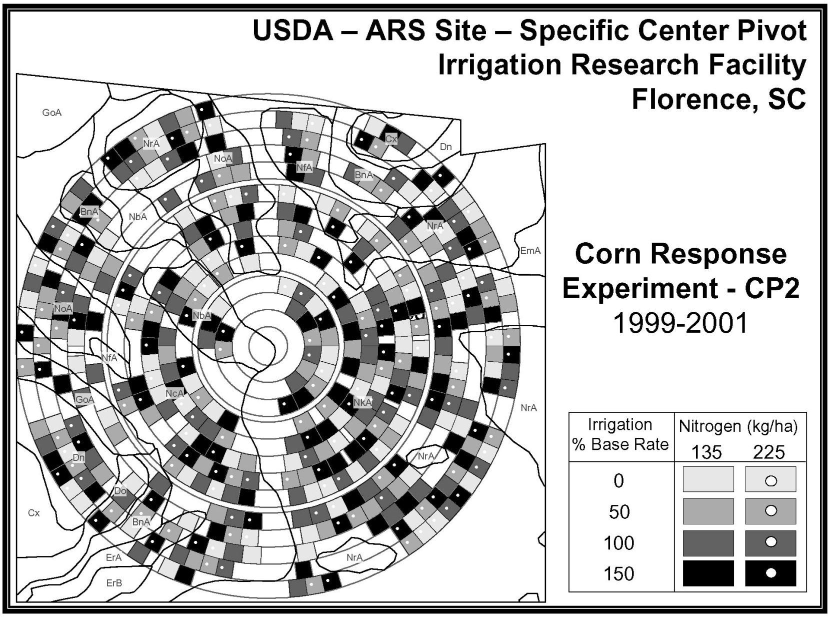


Figure 1. Diagram of experimental design for corn yield experiment, with soil map unit boundaries.

(USDA-SCS 1986); a brief description of the 12 soil map units under the center pivot can be found in Table 1 of Sadler et al. (2002). The design imposed in the experiment consisted of treatments of nitrogen and irrigation. Where sufficient area existed in the soil map unit boundaries a 4 × 2 factorial randomized complete block design (RCB) was introduced, while in regions where insufficient area was available randomized incomplete blocks (RICB) were used. This produced 39 RCB's and 19 RICB's for a total of 396 plots. The four irrigation treatments were 0%, 50%, 100%, and 150% of an irrigation base rate (IBR) determined using meteorological conditions combined with soil water potential values (SWP). The two N-fertilizer treatments were 135 kg/ha and 225 kg/ha, the recommended rainfed and irrigated rates. Moreover, these treatments corresponded to target yields of 6.3 and 10.1 Mg/ha. A diagram of the experimental design can be found in Figure 1 and a detailed description of the experiment is provided in Sadler et al. (2002).

As part of an analysis of variance Sadler et al. (2002) determined that corn yield can be described as a quadratic response curve in irrigation, that is,

$$\text{Yield} = A_0 + A_1\text{Irr} + A_2\text{Irr}^2 + \varepsilon, \tag{2.1}$$

where Irr is irrigation applied in mm. The problem with this function is that it does not take into account spatial variation or variation due to the imposed nitrogen treatment. However, as previously noted, this response curve provides important economic and managerial information to the practitioner and thus improving its overall utility is of general interest. For example, two important quantities that result from this function are rainfed yield (yield at 0 irrigation) and maximum yield; see Figure 2 for a schematic of useful quantities that can be derived from this curve.

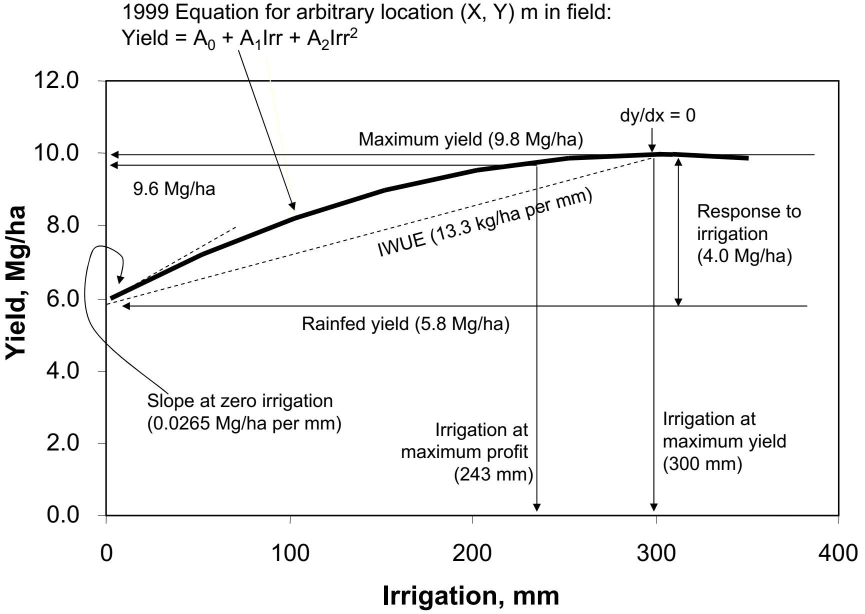


Figure 2. Example schematic diagram of managerial and economic quantities that can be derived from the quadratic response curve.

One approach that was proposed by Sadler (2004) to accommodate the inherent spatial variation was to use a “two-step sequential” method. The first part of the method isolated layers by treatment and then interpolated values to the plot centers for each of the 396 points for each level of the treatment. This procedure splits the experiment into eight unique treatments, four for the low level of nitrogen and four for the high level. Note that, by construction, this procedure ignores the relationship between N levels. Finally, the four values are extracted at each location and a regression run to produce estimates of the coefficients in (2.1). This produces 396 equations for each level of N (i.e., 792 equations total).

There are several distinct disadvantages that arise from the “two-step” procedure. First, there are 792 separate equations that the practitioner needs to evaluate in order to produce any desired quantity of interest. Second, the method does not use spatial correlation or “closely related” treatments in formulating parameter estimates. Further, the method requires that the practitioner form a kriging estimate in the first stage of the procedure (i.e. perform spatial interpolation). This presupposes explicit knowledge of spatial statistics and requires choosing several “tuning” parameters. Of course, the method we propose is not entirely free of these choices either. However, the practitioner is relieved from making these choices through the implementation of default knot choices (location and number) as well as through maximum likelihood (or REML) estimation of the smoothing parameter. Finally, and perhaps most importantly, there are no measures of uncertainty for any of the estimated quantities associated with the model or derived as a result.

The method proposed here addresses and solves each of the potential disadvantages associated with the "two-step" procedure. First, the model coefficients are spatially and nitrogen (treatment) varying resulting in model estimates obtained using all of the observations and hence "borrows strength" across the spatial and treatment domains. That is, all of the observations are used to estimate our model alleviating the need for 792 separate equations. Moreover, we provide pointwise confidence surfaces for model coefficients as well as describe methods for obtaining measures of uncertainty for all researcher-derived quantities of interest.

3. BIVARIATE PENALIZED SPLINE REGRESSION

In order to fully account for the spatial dependency present in our experiment requires a model capable of accounting for a continuous interaction between both directions in the spatial domain. To this end, the model ultimately proposed involves a bivariate predictor; therefore, we provide a brief overview of general spline based nonparametric regression for bivariate predictors. The description we provide closely follows that of Ruppert et al. (2003, chap. 13) and makes use of the equivalence between P-splines and mixed models. The first model we consider is the general bivariate smoothing model

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \tag{3.1}$$

where $\mathbf{x}_i = (s_i, t_i)$ can be thought of as *latitude* and *longitude* (or positions in x-direction and y-direction on a Cartesian coordinate system), respectively, and ε_i are assumed to be iid $\mathcal{N}(0, \sigma_\varepsilon^2)$. Specifically, for $\mathbf{x}_i$, $\boldsymbol{\kappa}_k \in \mathbb{R}^2$ let

$$\mathbf{X} = [1 \ \mathbf{x}_i]_{1\le i\le n}\,; \quad \mathbf{Z}_K = [C(\mathbf{x}_i - \boldsymbol{\kappa}_k)]_{\substack{1\le i\le n,\\ 1\le k\le K}}$$

where

$$C(\mathbf{r}) = ||\mathbf{r}||^{2\nu-2}\log||\mathbf{r}||, \tag{3.2}$$

and $\boldsymbol{\kappa}_k$ $(k = 1, \dots, K)$ denote fixed knot points in $\mathbb{R}^2$. In this context ν is an integer greater than one that controls smoothness. Additionally, note that $C(\mathbf{r})$ can also be chosen as radial basis functions corresponding to a proper covariance structure (see Ruppert et al. 2003, p. 254). Further, let

$$\boldsymbol{\Omega}_K = [C(\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'})]_{1\le k,k'\le K},$$

then the penalized spline regression is obtained by minimizing

$$\frac{1}{\sigma_\varepsilon^2}||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}_K\mathbf{b}||^2 + \frac{1}{\lambda\sigma_\varepsilon^2}\mathbf{b}'\boldsymbol{\Omega}_K\mathbf{b},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, $\mathbf{b} = (b_1, \dots, b_K)'$ and λ corresponds to a fixed penalty parameter ("smoothing" parameter). In addition, take $\boldsymbol{\beta}$ to be fixed, $\mathbf{b}$ random with $\mathrm{E}(\mathbf{b}) = \mathbf{0}$, $\mathrm{cov}(\mathbf{b}) = \sigma_u^2\boldsymbol{\Omega}_K^{-1}$, where $\sigma_u^2 = \lambda\sigma_\varepsilon^2$. So long as $(\mathbf{b}', \boldsymbol{\varepsilon}')'$ is normally distributed, where

$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ and $\mathbf{b}$ and $\boldsymbol{\varepsilon}$ are independent one can obtain an equivalent linear mixed model representation of the penalized spline (Brumback, Ruppert, and Wand 1999); see Crainiceanu, Ruppert, and Wand (2005) and references therein for complete details. That is, the P-spline is equal to the best linear unbiased predictor (BLUP) in the linear mixed model (LMM)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_K\mathbf{b} + \boldsymbol{\varepsilon}, \tag{3.3}$$

with

$$\operatorname{cov}\begin{pmatrix} \mathbf{b} \\ \boldsymbol{\varepsilon} \end{pmatrix} = \begin{pmatrix} \sigma_u^2 \boldsymbol{\Omega}_K^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma_\varepsilon^2 \mathbf{I}_n \end{pmatrix}.$$

Again, following Crainiceanu et al. (2005), define $\mathbf{Z} = \mathbf{Z}_K \boldsymbol{\Omega}_K^{-1/2}$ and $\mathbf{u} = \boldsymbol{\Omega}_K^{1/2}\mathbf{b}$. Then the mixed model (3.3) can be equivalently expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{3.4}$$

with

$$\operatorname{cov}\begin{pmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{pmatrix} = \begin{pmatrix} \sigma_u^2 \mathbf{I}_K & \mathbf{0} \\ \mathbf{0} & \sigma_\varepsilon^2 \mathbf{I}_n \end{pmatrix}$$

and fit using best linear unbiased predictor (BLUP) estimation. Finally, if K (the number of knot points) is large, the function $f(\cdot)$ in (3.1) can usually be approximated sufficiently well.

Perhaps the most difficult aspect with regard to bivariate semiparametric regression is the choice of knots. In the univariate case one reasonable option for selecting knot points is to evenly distribute them in the quantile domain (French, Kammann, and Wand 2001) according to rules such as

$$K = \min\left\{\frac{1}{4}(\text{number of unique predictor values}), 35\right\}, \tag{3.5}$$

or those given in Ruppert (2002); see Ruppert et al. (2003) for a complete discussion. Unfortunately this approach breaks down in higher dimensions due to the lack of a clear cut definition of the quantile (for $d > 1$). One method used to remedy this situation is to place knot points using *space filling* designs (Johnson, Moore, and Ylvisaker 1990; Nychka and Saltzman 1998). However, the use of space filling designs can be slow for large n and K; therefore, as suggested by Ruppert et al. (2003, chap. 13, pp. 255–260), we apply a space filling design algorithm to a set of randomly selected points $\mathbf{x}_i$. This can be achieved using the `FUNFITS` module (Nychka, Haaland, O'Connell, and Ellner 1998) or via the SemiPar 1.0 package in R (Wand et al. 2005).

In order to simplify matters, in terms of choosing user-defined "tuning" parameters, we implement a slight adaptation of the default bivariate smoothing algorithm of Ruppert et al. (2003, p. 257). The algorithm we implement is a slight adaptation only to the extent that the procedure suggested by Ruppert et al. (2003) is in terms of a simple bivariate smoothing (i.e., only one bivariate predictor—namely, location); however, in our case the algorithm is applied to a spatially treatment varying coefficient model having several predictors (both univariate and bivariate). Specifically, for our application, the algorithm we implement chooses the number of knots according to $K = \max(20, \min(n/4, 150))$, and obtains the appropriate knot points by applying a space filling algorithm. Note that this step can also be executed using the default knot selection algorithm in the SemiPar 1.0 R package (Wand et al. 2005). Further, we choose $\nu = 2$ in (3.2) and reexpress the model in the form of (3.4). Then using standard software we fit the model with $\sigma_\varepsilon^2, \sigma_u^2$ chosen by maximum likelihood and $\boldsymbol{\beta}$, $\mathbf{u}$ the corresponding EBLUPs (Ngo and Wand 2004).

4. SPATIALLY TREATMENT VARYING RESPONSE CURVES

One of the goals of all experiments is to describe the response to treatment factors, as in Kuehl (2000). That is, when treatment factors have qualitative levels it is often beneficial to characterize the response y to the factor levels x using polynomial regression, sometimes referred to as response curves. Furthermore, response curves have the advantage of providing a graphical representation in which one can visualize the response across the different factor levels of the treatment included in the experiment; see Kuehl (2000) and the references therein for further details.

Many times when conducting agricultural experiments the response varies not only across the factor levels of the treatment but across the spatial domain as well. Therefore making the response curve spatially explicit will potentially provide better explanatory power by borrowing strength among spatially correlated responses. In particular, suppose that our response can be described as a quadratic function of the treatment (trt). Then

$$Y_i = \beta_0(\mathbf{x}_i) + \beta_1(\mathbf{x}_i)\mathrm{trt}_i + \beta_2(\mathbf{x}_i)\mathrm{trt}_i^2 + \varepsilon_i \tag{4.1}$$

provides a spatially varying response curve. Furthermore, if there are known treatment A by treatment B interactions they can be implicitly incorporated into (4.1) yielding

$$Y_i = \beta_0(\mathbf{x}_i, \mathrm{trt}A_i) + \beta_1(\mathbf{x}_i, \mathrm{trt}A_i)\mathrm{trt}B_i + \beta_2(\mathbf{x}_i, \mathrm{trt}A_i)\mathrm{trt}B_i^2 + \varepsilon_i. \tag{4.2}$$

This method of modeling response curves is quite flexible as there are numerous ways one can express the smooth surfaces (or curves) $\beta_0(\cdot,\cdot)$, $\beta_1(\cdot,\cdot)$, and $\beta_2(\cdot,\cdot)$. Although these coefficient surfaces can be modeled semiparametrically via kernels (i.e., local polynomials), Fourier bases, wavelets, etc., in keeping with our previous exposition and with the goal of producing "default" spatially treatment varying response curves we use the mixed model formulation of the P-spline.

Before considering models with bivariate radial basis functions, we examined several models of the form

$$Y_i = A_0(s_i, t_i, N_i) + A_1(s_i, t_i, N_i)\mathrm{Irr}_i + A_2(s_i, t_i, N_i)\mathrm{Irr}_i^2 + \varepsilon_i. \tag{4.3}$$

Specifically, we considered additive models for $A_\bullet(s_i, t_i, N_i)$ with a binary offset (in Nitrogen) (cf. Ruppert et al. 2003, pp. 162–163). Under this specification, the general form of (4.3) can then be expressed as

$$\begin{aligned} Y_i &= \{a_0 N_i + g_0(s_i) + h_0(t_i)\} + \{a_1 N_i + g_1(s_i) + h_1(t_i)\}\mathrm{Irr}_i \\ &\quad + \{a_2 N_i + g_2(s_i) + h_2(t_i)\}\mathrm{Irr}_i^2 + \varepsilon_i, \end{aligned} \tag{4.4}$$

and, with an appropriately chosen set of univariate basis functions (i.e., low-rank, thin-plate splines), can subsequently written in the form of (3.4).

Ultimately, due to the spatial attributes of our experiment (and model selection criteria), the model used in our analysis had the general form

$$Y_i = A_0(\mathbf{x}_i, N_i) + A_1(\mathbf{x}_i, N_i)\mathrm{Irr}_i + A_2(\mathbf{x}_i, N_i)\mathrm{Irr}_i^2 + \varepsilon_i, \tag{4.5}$$

where $\mathbf{x}_i = (s_i, t_i)$, Irr_i, and N_i refer to the spatial location, the irrigation treatment and the nitrogen treatment, respectively, for the ith observation. The specific form of (4.5) we constructed consists of a bivariate smoother for $A_\bullet(\mathbf{x}_i, N_i)$ with a binary offset (in Nitrogen). To this end, (4.5) can be expressed as

$$Y_i = \{a_0 N_i + f_0(\mathbf{x}_i)\} + \{a_1 N_i + f_1(\mathbf{x}_i)\}\mathrm{Irr}_i + \{a_2 N_i + f_2(\mathbf{x}_i)\}\mathrm{Irr}_i^2 + \varepsilon_i, \tag{4.6}$$

where the specific form of $f_\bullet(\mathbf{x}_i)$ is discussed in Section 3. Additionally, define

$$\begin{aligned} \mathbf{X} &= \left[N_i \quad \mathbf{x}_i \quad N_i\mathrm{Irr}_i \quad \mathbf{x}_i\mathrm{Irr}_i \quad N_i\mathrm{Irr}_i^2 \quad \mathbf{x}_i\mathrm{Irr}_i^2 \right]_{1 \le i \le n}, \\ \mathbf{Z}_K &= \underset{1 \le k \le K}{[||\mathbf{x}_i - \boldsymbol{\kappa}_k||^2 \log||\mathbf{x}_i - \boldsymbol{\kappa}_k||]_{1 \le i \le n}}, \end{aligned}$$

and

$$\boldsymbol{\Omega}_K = \underset{1 \le k, k' \le K}{[||\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'}||^2 \log||\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'}||]}.$$

Next, let $\mathbf{Z}_{st} = \mathbf{Z}_K \boldsymbol{\Omega}_K^{-1/2}$, $\mathbf{Z}_0 = \mathbf{Z}_{st}$, $\mathbf{Z}_1 = \mathbf{Z}_{st} * \mathbf{Irr}$, and $\mathbf{Z}_2 = \mathbf{Z}_{st} * \mathbf{Irr}^2$ where $*$ denotes column multiplication that is,

$$\begin{pmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{pmatrix} * \begin{pmatrix} \mathrm{Irr}_1 \\ \mathrm{Irr}_2 \end{pmatrix} = \begin{pmatrix} z_{11}\mathrm{Irr}_1 & z_{12}\mathrm{Irr}_1 \\ z_{21}\mathrm{Irr}_2 & z_{22}\mathrm{Irr}_2 \end{pmatrix}.$$

Then denote $\mathbf{Z}^* = [\mathbf{Z}_0 \quad \mathbf{Z}_1 \quad \mathbf{Z}_2]$. Thus it is straightforward to rewrite (4.6) in the form of (3.4), with $\mathbf{Z}$ replaced by $\mathbf{Z}^*$, where it can be easily estimated using standard software.

As previously discussed, in order to analyze the data at hand several competing models were considered. Specifically, we considered various models of the form (4.4) along with our “default” bivariate model (4.6). One approach we used in arriving at a final model was to compare model selection criteria for competing models. In particular, we used AIC; in this context AIC is defined by (see Simonoff and Tsai 1999)

$$\mathrm{AIC} \equiv \log(\mathrm{RSS}_\lambda) + \frac{df_{\mathrm{fit},\lambda}}{n}, \tag{4.7}$$

where

$$\mathrm{RSS}_\lambda = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

Note that $df_{\text{fit},\lambda} = \text{tr}(\mathbf{S}_\lambda)$, where

$$\mathbf{S}_\lambda = \mathbf{C}(\mathbf{C}'\mathbf{C} + \Lambda)^{-1}\mathbf{C}',$$

is the *smoother* matrix, $\mathbf{C} \equiv [\mathbf{X}|\mathbf{Z}]$ and

$$\Lambda \equiv \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_\varepsilon^2 \text{cov}(\mathbf{u})^{-1} \end{pmatrix} \tag{4.8}$$

(see Ruppert et al. 2003, p. 175). Although using AIC, as well as other information-based model selection criteria, has become somewhat common when comparing mixed models differing in their fixed effects, the appropriateness of such comparisons is questionable if the models are estimated via REML (Verbeke and Molenberghs 2000). In fact, these comparisons are incorrect as can be seen by considering the derivation of AIC through Kullback–Leibler distances (Burnham and Anderson 2002). As a result of this limitation we estimate all of the models using maximum likelihood. Although we incorporate the use of AIC in our model selection procedure, ultimately we choose our model based on a residual analysis and on scientific knowledge of the underlying experiment.

Specifically, the exploratory analysis we conducted coupled with scientific knowledge regarding the specific experiment under consideration helped simplify (4.4) by assuming $h_\bullet(\cdot)$ to be linear in t. Furthermore, the model that resulted from this assumption, using low-rank (cubic) thin-plate splines, was deemed best among several competing models investigated from this class of models using AIC. Subsequently this model was compared with (4.6) and it was determined that the bivariate spatial model was superior. In addition to having a smaller AIC value (6.44 versus 6.70) the model was favored on the merit of several other measures. First the coefficient of determination (R^2) was highest for the bivariate model (0.72 versus next highest 0.61). More importantly the residuals from this model appeared to be normally distributed (p-value for Shapiro–Wilks test $= 0.14$; along with visual inspection) and uncorrelated while the other models yielded skewed residuals (p-values for Shapiro-Wilks test all less than 0.01). In summary, the preferred model was the "default" bivariate spatially nitrogen varying response curve.

One important aspect associated with any modeling endeavor is to attach measures of uncertainty to the estimated quantities. By taking advantage of the mixed model formulation of the P-spline these measures are straight forward to calculate. That is, letting

$$M_i = \text{diag}_i \left\{ \mathbf{C}(\mathbf{C}'\mathbf{C} + \Lambda)^{-1}\mathbf{C}' \right\},$$

a pointwise confidence surface (with bias allowance) for yield is given by

$$\widehat{Y}_i \pm z_{(1-\alpha/2)} \widehat{\sigma}_\varepsilon^2 \sqrt{M_i}. \tag{4.9}$$

It should be noted that because in our analysis we have a large sample size ($n = 396$; although the effective sample size may be smaller) we use the standard normal when calculating the (pointwise) confidence surface. However, in smaller sample sizes we can replace the standard normal with the appropriate t-distribution (cf. Ruppert et al. 2003, p. 159).

As depicted in Figure 2, several quantities are of interest to the practitioner in addition to yield. One quantity of general interest is the rainfed yield, $A_0(\mathbf{x}_i, N_i)$. However, without measures of uncertainty this quantity is of diminished usefulness. In general, uncertainty measures for this, or any, coefficient surface can be easily handled. To construct uncertainty measures let P denote the number of columns in $\mathbf{C}$ and $\{\mathcal{I}_0, \ldots, \mathcal{I}_d\}$ be a partition of the column indices in $\mathbf{C}$ such that $\mathcal{I}_0$ corresponds to β_0 and $\mathcal{I}_j$ corresponds to $f_j(\cdot)$ for $j = 1, \ldots, d$ (see Ruppert et al. 2003, p. 175). Further, define $\mathbf{E}_m$ for $m = 0, 1, 2$ to be the $P \times P$ matrix with ones in the diagonals corresponding to the partition of the columns of $\mathbf{C}$ associated with the coefficient surface of interest A_m and

$$ME_{mi} = \text{diag}_i \left\{ \mathbf{C}\mathbf{E}_m(\mathbf{C}'\mathbf{C} + \Lambda)^{-1}\mathbf{C}' \right\}.$$

Then a pointwise confidence surface is given by

$$\widehat{A}_m(\mathbf{x}_i, N_i) \pm z_{(1-\alpha/2)}\widehat{\sigma}_\varepsilon^2 \sqrt{ME_{mi}} \tag{4.10}$$

(see Ruppert et al. 2003, p. 175). It is often the case that the practitioner will be interested in even more complicated quantities. For instance, the maximum response to irrigation is defined to be the maximum within the imposed treatment range. For concave down forms ($A_2(\mathbf{x}_i, N_i) < 0$), this can be found by evaluating (4.5) at the point where the derivative is equal to zero. This value of irrigation can be found several different ways, one being to estimate the derivative function directly. Then, using obvious notation,

$$\widehat{\text{Irr}}_{i,\max} = \frac{-\widehat{A}_1(\mathbf{x}_i, N_i)}{2\widehat{A}_2(\mathbf{x}_i, N_i)} \tag{4.11}$$

defines the level of irrigation producing the maximum yield for a specific location $\mathbf{x}_i$ and a given nitrogen treatment. Note that choosing maximums in this manner rather than selecting the empirical maximum from the experiment allows intermediate levels of irrigation to be selected as a hypothetical maximum. Additionally, uncertainty measures for this and other complex quantities of interest can be found using the Delta method (see Bickel and Doksum 2001) or by using the *mixed model bootstrap* (Kaurmann, Claeskens, and Opsomer 2009). Furthermore, predicting yield for different treatment combinations at a specific spatial location in the field, for instance to determine optimal treatment combinations, can be facilitated by "plugging in" the appropriate values to the estimated curve. Finally, appropriate pointwise prediction surfaces can be constructed by

$$\widehat{Y}_j \pm z_{(1-\alpha/2)}\widehat{\sigma}_\varepsilon^2 \sqrt{1 + ||\ell_j||^2}, \tag{4.12}$$

where

$$||\ell_j|| = \sqrt{\mathbf{C}_j(\mathbf{C}'\mathbf{C} + \Lambda)^{-1}\mathbf{C}'\mathbf{C}(\mathbf{C}'\mathbf{C} + \Lambda)^{-1}\mathbf{C}_j'},$$

and $\mathbf{C}_j$ corresponds to the values associated with a new observation y_j (see Ruppert et al. 2003, p. 138).

The original goals of the analysis undertaken in Sadler et al. (2002) were to evaluate the mean response of corn to irrigation amounts on 12 soil map units and to compare the variation in the response within and among soil map units. Combining Figures 1 and 3 provides

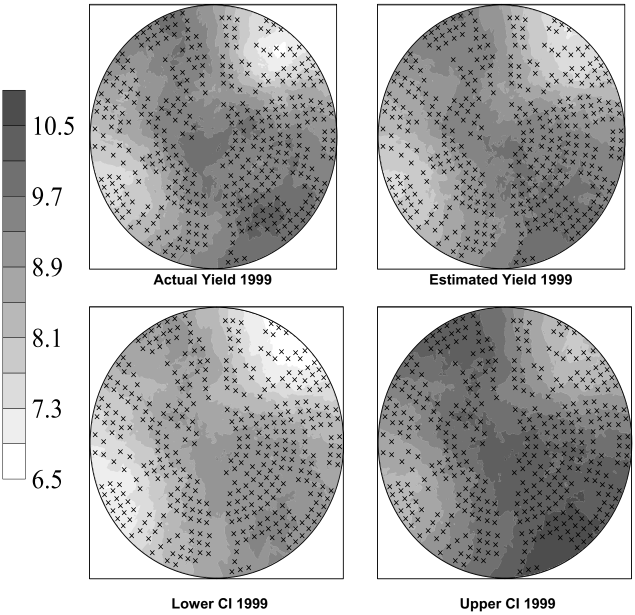


Figure 3. Yield surfaces for 1999, including estimated yield along with upper and lower 95% (pointwise) confidence surfaces.

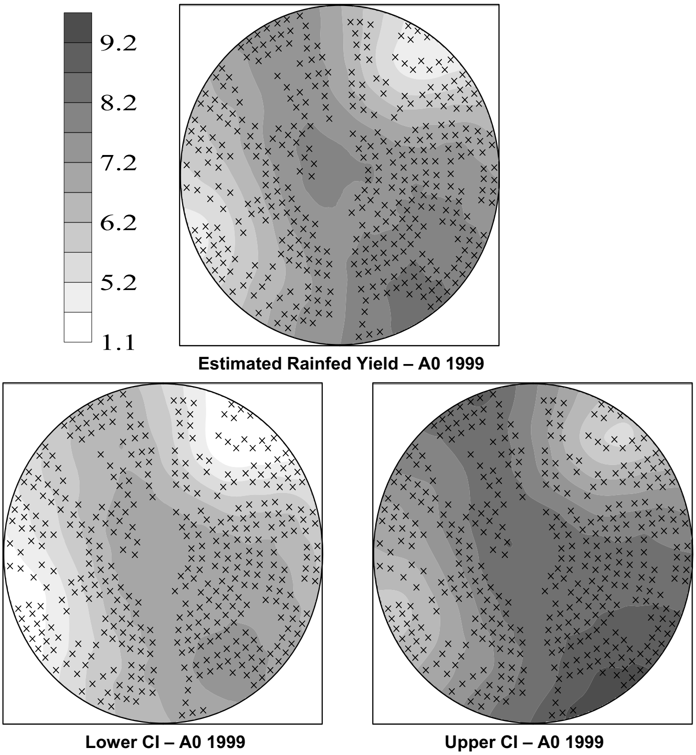
a very lucid way of delivering this information graphically. As a result of these graphical displays it can be determined that the response to irrigation is significantly different between the soil map units Cx (Coxville loam) and NkA (Norfolk lfs). However, for NkA and NbA (Noboco lfs) there is no significant difference (see Figure 3). Similar types of statements can be made regarding the response to rainfed yield, see Figure 4. While formal economic and managerial implications are beyond the scope of this article, the methods detailed here can be used to conduct such an analysis.

5. BAYESIAN IMPLEMENTATION

The method we propose can also be implemented in a Bayesian framework. In order to carry out a Bayesian analysis we take advantage of the mixed model formulation in (3.4). Following Ruppert et al. (2003), let $\mathbf{u} = (\mathbf{u_0}', \mathbf{u_1}', \mathbf{u_2}')'$ corresponding to the partition of $\mathbf{Z}^*$; the mixed model formulation specifies a $\mathcal{N}(0, \sigma_{u_i}^2\mathbf{I})$ prior on $\mathbf{u}_i$ $(i = 0, 1, 2)$ as well as the likelihood

$$[\mathbf{y}|\boldsymbol{\beta},\ \mathbf{u},\ \sigma_{u_0}^2,\ \sigma_{u_1}^2,\ \sigma_{u_2}^2,\ \sigma_\varepsilon^2].$$

To completely specify a Bayesian model requires us to choose priors for $(\boldsymbol{\beta},\ \sigma_{u_0}^2,\ \sigma_{u_1}^2,\ \sigma_{u_2}^2,\ \sigma_\varepsilon^2)$. Since little information is known about $\boldsymbol{\beta}$ we impose an improper uniform prior, that is, $[\boldsymbol{\beta}] \equiv 1$. Although we chose an improper uniform prior for $\boldsymbol{\beta}$ one could choose the proper prior $\mathcal{N}(0, \sigma_\beta^2\mathbf{I})$ where σ_β^2 is large. This choice would produce a proper prior which is essentially uniform over the range of $\boldsymbol{\beta}$. Additionally, we chose inverse gamma priors



Estimated Rainfed Yield – A0 1999

Lower CI – A0 1999

Upper CI – A0 1999

Figure 4. Estimated rainfed yield surface for 1999 (yield at irrigation = 0; i.e. A_0), including upper and lower 95% (pointwise) confidence surfaces.

for $\sigma^2_{u_i}$ $(i = 0, 1, 2)$ and σ^2_ε. That is, for $i = 0, 1, 2$

$$\begin{aligned}\sigma^2_{u_i} &\sim \mathrm{IG}(A_{u_i},\ B_{u_i}),\\ \sigma^2_\varepsilon &\sim \mathrm{IG}(A_\varepsilon,\ B_\varepsilon),\end{aligned}$$

where $A_{u_i},\ B_{u_i},\ A_\varepsilon,\ B_\varepsilon$ are chosen equal to 0.1 and thus yield a noninformative but proper prior (Ruppert et al. 2003, p. 280). The model is constructed as a hierarchical Bayes model and invoking conditional independence properties the posterior distribution is given by

$$\begin{aligned}[\boldsymbol{\beta},\ \mathbf{u},\ \sigma^2_{u0},\ \sigma^2_{u_1},\ \sigma^2_{u_2},\ \sigma^2_\varepsilon|\mathbf{y}] \propto{}& [\mathbf{y}|\boldsymbol{\beta},\ \mathbf{u},\ \sigma^2_\varepsilon][\mathbf{u_0}|\sigma^2_{u0}][\mathbf{u_1}|\sigma^2_{u_1}]\\ &\times[\mathbf{u_2}|\sigma^2_{u_2}][\sigma^2_{u0}][\sigma^2_{u_1}][\sigma^2_{u_2}][\boldsymbol{\beta}][\sigma^2_\varepsilon].\end{aligned}$$

Let $\boldsymbol{\Lambda}$ be defined as in (4.8). Then the full conditional of $(\boldsymbol{\beta},\ \mathbf{u})$ is

$$[\boldsymbol{\beta},\ \mathbf{u}|\sigma^2_\varepsilon,\ \sigma^2_{u0},\ \sigma^2_{u_1},\ \sigma^2_{u_2},\ \mathbf{y}] \sim \mathcal{N}\left\{(\mathbf{C}'\mathbf{C}+\boldsymbol{\Lambda})^{-1}\mathbf{C}'\mathbf{y},\ \sigma^2_\epsilon(\mathbf{C}'\mathbf{C}+\boldsymbol{\Lambda})^{-1}\right\}.$$

Further for $i = 0, 1, 2$

$$[\sigma^2_{u_i}|\mathbf{y},\ \boldsymbol{\beta},\ \mathbf{u}_i] \sim \mathrm{IG}\left(A_{u_i} + \frac{1}{2}K, B_{u_i} + \frac{1}{2}||\mathbf{u}_i||^2\right),$$

where K denotes the number of knots. Finally the full conditional for σ_{ϵ}^2 is

$$[\sigma_{\varepsilon}^2 | \mathbf{y},\ \boldsymbol{\beta},\ \mathbf{u}_i\ \sigma_{u_0}^2,\ \sigma_{u_1}^2,\ \sigma_{u_2}^2] \sim \mathrm{IG}\left(A_{\varepsilon} + \frac{1}{2}n, B_{\varepsilon} + \frac{1}{2}||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}||^2\right).$$

Estimation of the model is then achieved using Gibbs sampling (e.g., Casella and George 1992). Although, for analysis, we implemented Gibbs sampling directly, it can also be carried out using the software WinBUGS (Spiegelhalter et al. 2003; Crainiceanu et al. 2005).

We conducted a Bayesian analysis using a stand alone program implementing Gibbs sampling. Specifically, we used 45,000 MCMC iterations and discarded the first 10,000 for burn-in. The results of the analysis corroborated the results found in Section 4 and therefore are not presented here; rather they are available on request.

6. DISCUSSION

Studies of the effects of irrigation nonuniformity on the crop response comprise an important topic in agricultural experiments. Spatial variation in crop response to water naturally leads to the topic of variable-rate or site-specific agriculture. However, most often the reported values are means across spatial replications. Further, crop response functions have historically been determined under uniform irrigation management conditions. Thus there has been widespread interest in whether relationships obtained under uniform management can be applied in site-specific agriculture. Finally properly constructed response functions can be used to provide essential economic and managerial information that allow the practitioner to make strategic decision about profitability.

In order to address these questions we develop spatially treatment varying coefficient models. The models we propose are quite flexible and can be adapted in a straightforward manner for use in a broad class of agricultural problems. Additionally, by taking advantage of the mixed model representation of the P-spline (Brumback, Ruppert, and Wand 1999) we are able to express our models in a form that can be easily fit using standard mixed model software. In addition, the mixed model representation facilitates easy calculation of uncertainty measures either analytically or through mixed model bootstrap methodology (Kaurmann, Claeskens, and Opsomer 2009). Furthermore, we provide a Bayesian version of the model that is equally easy to implement. Computer code is available upon request from the first author for all analyses conducted in this article. Finally, we have demonstrated the effectiveness of our approach through the analysis of site-specific agricultural data from the USDA-ARS.

ACKNOWLEDGMENTS

Mention of specific companies, products, or trade names is made only to provide information to the reader and does not constitute endorsement by the USDA-Agricultural Research Service. Additionally, the authors would like to thank Scott Drummond for his assistance in preparing the figures. Finally, the authors would also like to thank two anonymous referees whose suggestions helped improve this article.

[Received February 2007. Revised January 2008.]

REFERENCES

Agarwal, D. K., Gelfand, A. E., Sirmans, C. F. and Thibadeau, T. G. (2003), "Flexible Nonstationary Spatial House Price Models," Technical Report, Department of Statistics, University of Connecticut.

Assunçao, R. M., Potter, J., and Cavenaghi, S. M. (2002), "A Bayesian Space Varying Parameter Model Applied to Estimating Fertility Schedules," *Statistics in Medicine*, 21, 2057–2075.

Assunçao, J. J., Gamerman, D., and Assunçao, R. M., (1999), "Regional Differences in Factor Productivities of Brazilian Agriculture: A Space-Varying Parameter Approach," Technical Report, Universidade Federal do Rio de Janeiro, Statistical Laboratory.

Bickel, P., and Doksum, K. (2001), *Mathematical Statistics,* New Jersey: Prentice Hall.

Banerjee, S., and Johnson, G. (2006), "Coregionalized Single- and Multiresolution Spatially Varying Growth Curve Modeling with Application to Weed Growth," *Biometrics*, 62, 864–876.

Brumback, B., Ruppert, D., and Wand, M. P. (1999), Comment on "Variable Selection and Function Estimation in Additive Nonparametric Regression Using Data-based Prior by Shively, Kohn and Wood," *Journal of the American Statistical Association*, 94, 794–797.

Burnham, K., and Anderson, D. (2002), *Model Selection And Multimodel Inference: A Practical Information-Theoretic Approach,* New York: Springer-Verlag.

Casella, G., and George, E. (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167–174.

Crainiceanu, C., Ruppert, D., and Wand, M. P. (2005), "Bayesian Analysis for Penalized Spline Regression Using WinBUGS," *Journal of Statistical Software*, 14.

Eilers, P. H. C. L., and Mark, B. D. (1996), "Flexible Smoothing with B-splines and Penalties" (with discussion), *Statistical Science*, 11, 89–121.

Fahrmeir, L., Kneib, T., and Lang, S. (2004), "Penalized Structured Additive Regression for Space-time Data: A Bayesian Perspective," *Statistica Sinica*, 14, 731–761.

French, J. L., Kammann, E. E., and Wand, M. P. (2001), Comment on Paper by Ke and Wang, *Journal of the American Statistical Association*, 96, 1285–1288.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2002), *Geographically Weighted Regression,* New York: John Wiley.

Gamerman, D., Moreira, A. R. B., and Rue, H. (2002), "Space-Varying Regression Models," Technical Report, Universidade Federal do Rio de Janeiro, Statistical Laboratory.

Gelfand, A., Kim, H. J., Sirmans,C. F., and Banerjee, S. (2003), "Spatial Modeling with Spatially Varying Coefficient Processes," *Journal of the American Statistical Association*, 98, 387–396.

Gimenez, O., Barbraud, C., Crainiceanu, S., Jenouvrier, S., and Morgan, B. J. T. (2006), "Semiparametric Regression in Capture-Recapture Modelling," *Biometrics*, 62, 691–698.

Hastie, T. J., and Tibshirani, R. J. (1993), "Varying-coefficient Models," *Journal of the Royal Statistical Society*, Series B, 55, 757–796.

Johnson, M. E., Moore, L. M., Ylvisaker, D. (1990), "Minimax and Maximum Distance Designs," *Journal of Statistical Planning and Inference*, 26, 131–148.

Kaurmann, G., Claeskens, G., and Opsomer, J. D. (2009), "Bootstrapping for Penalized Spline Regression," *Journal of Computational and Graphical Statistics*, to appear.

Kuehl, R. (2000), *Design of Experiments: Statistical Principles of Research Design and Analysis* (2nd ed.), Pacific Grove, CA: Duxbury Press.

Luo, Z., and Wahba, G. (1998), "Spatio-Temporal Analogues of Temperature Using Smoothing Spline ANOVA," *Journal of Climatology*, 11, 18–28.

Ngo, L., and Wand, M. P. (2004), "Smoothing with Mixed Model Software," *Journal of Statistical Software*, 9.

Nychka, D., Haaland, P., O'Connell, M., and Ellner, S. (1998), "FUNFITS, Data Analysis and Statistical Tools for Estimating Functions," in *Case Studies in Environmental Statistics*, Lecture Notes in Statistics, Vol. 32, eds. D. Nychka, W.W. Piegorsch, and L.H. Cox, New York: Springer-Verlag, 159–179.

Nychka, D., and Saltzman, N. (1998), "Design of Air Quality Monitoring Networks," in *Case Studies in Environmental Statistics*, Lecture Notes in Statistics, Vol. 32, eds. D. Nychka, W.W. Piegorsch, and L.H. Cox, New York: Springer-Verlag, 51–76.

Ruppert, D. (2002), "Selecting the Number of Knots for Penalized Splines," *Journal of Computational and Graphical Statistics*, 11, 735–757.

Ruppert, D., Wand, M., and Carroll, R. (2003), *Semiparametric Regression,* Cambridge: Cambridge University Press.

Sadler, E. J. (2004), "Comparison of Four Analyses of Spatial Irrigation Production Functions," *NCR180 Meeting*, Conference Presentation.

Sadler, E. J., Camp, C. R., Evans, D. E., and Millen, J. A. (2002), "Spatial Variation of Corn Response to Irrigation," *Transactions of the ASAE*, 45, 1869–1881.

Simonoff, J., and Tsai, C. L. (1999), "Semiparametric and Additive Model Selection Using and Improved Akaike Information Criterion," *Journal of Computational and Graphical Statistics*, 8, 22–40.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003), *WinBUGS User Manual. Version 1.4* (*http://www.mrcbsu.cam.ac.uk/bugs*), Technical Report, Medical Research Council Biostatistics Unit, Cambridge, U.K.

USDA-SCS (1986), "Classification and Correlation of Soils of Coastal Plains Research Center," ARS, Florence, South Carolina. Ft. Worth, Texas: USDA-SCS South National Technical Center.

Verbeke, G., and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data,* New York: Springer-Verlag.

Wand M. P., Coull, B. A., French, J. L., Ganguli, B., Kammann, E. E., Staudenmayer, J., and Zanobetti, A. (2005), SemiPar 1.0 R package. *http://cran.r-project.org*.

Wikle, C., Berliner, L., and Cressie, N. (1998), "Hierarchical Bayesian Space-Time Models," *Environmental and Ecological Statistics*, 5, 117–154.

Wikle, C., and Anderson, C. (2003), "Climatological Analysis of Tornado Report Counts Using a Hierarchical Bayesian Spatio-Temporal Model," *Journal of Geophysical Research*, 108(D24), 9005, doi: 10.1029/2002JD002806.

Zhao, Y., Staudenmayer, J., Coull, B. A., and Wand, M. P. (2006), "General Design Bayesian Generalized Mixed Models," *Statistical Science*, 20, 35–51.