Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 68(2) 2014 257-272

A Fay-Herriot Type Approach for Better Prediction in Multi-indexed Response with Application to Arctic Seawater Data Analysis

Ujjal Mukherjee[1] **and Snigdhansu Chatterjee**[2]
[1]*Carlson School of Management, University of Minnesota, Minneapolis MN 55455*
[2]*School of Statistics, University of Minnesota, Minneapolis MN 55455*

Received 03 January 2014; Accepted 19 May 2014

SUMMARY

We consider the problem of fitting a nonparametric curve to Arctic Ocean temperature data. Since several alternative curves may be fitted, we consider borrowing strength over fitted curves to create an esemble fit. This lead to a novel exercise involving nonparametric curve fitting and small area methods. Our results indicate that climate data analysis is a complex process, and standard statistical techniques may need to be considerably enhanced for applicability to big data arising from climate studies.

Keywords: Small area, Fay-Herriot model, Local polynomial regression, Arctic Ocean temperature.

1. INTRODUCTION

In recent years there has been a growing awareness and concern about climate change and its effects, both among the general populace and experts in various areas related to climate study. Natural phenomena like climate patterns are complex, typically non-stationary, spatio-temporal processes involving several variables; the relationship between the climate variables, and among the parameters of the distribution of such variables are not fully understood at present.

For example, consider the dataset on global seawater (Schmidt *et al.* 1999), accessible from the repository (http://data.giss.nasa.gov) of the Godard Institute of Space Studies, which will be the focus for the rest of this paper. This dataset includes measurements of temperature, salinity, deuterium, the ratio of the O-18 and O-16 isotopes of oxygen; co-variable information about the depth of the sea, the latitude and longitude, and the month and year at which the data was collected; along with references and notes. It is a compilation of data gathered by various teams of researchers at different points of time and location. Calibrations are carried out to correct for the difference in standards, techniques and instruments used by these teams, and such corrections are flagged. Missing values are present. Further technical minutiae relating to the dataset is available in the aforementioned website.

The pattern of seawater temperature, at any given latitude and longitude and depth of ocean, during a given season, is complex. We may study this pattern assuming that the seawater temperature is a random variable. The parameters or processes governing the seawater temperature distribution possibly depends on other variables like the salinity, North Atlantic or Arctic Oscillations, solar activity, and anthropogenic factors.

Corresponding author: Snigdhansu Chatterjee
E-mail address: chatt019@umn.edu

In order to study such patterns, there are typically two widely used approaches. First, we may attempt parametric distributional modeling, which typically involves making simplifying assumptions and restricting the class of features that are allowable in the data. For example, we may assume that some transformation of the seawater temperature is Normally distributed with a mean that is a linear combination of the other climate variables, and constant variance. Moreover, we might assume that conditional on other climate variables, seawater temperature at various locations and time points are independently distributed. Elementary exploratory analysis of data suggests such models are overly simple and do not explain most properties of the observed data, however, similar models are routinely used in climate data analysis (Allen and Tett 1999, Hasler *et al.* 2009). More systematic analysis, but still keeping to the parametric statistical paradigm, might involve including parametric spatio-temporal dependency patterns, heteroscedasticity, and hierarchical modeling assumptions. These are quite common in Bayesian modeling of climate and other kinds of data, which is also practiced by a large group of researchers (Lee *et al*. 2005, Min *et al*. 2005, Bhattacharjee and Chatterjee 2013). For example, we may use a spatio-temporal conditional autoregressive model to build-in space-time dependency, and allow the hyper-parameters of this conditional autoregression to vary to reflect non-stationarity.

Bayesian, or other complex parametric modeling, of climate data, is satisfactory in many cases, but not in all situations. To start with, complex parametric models might obfuscate the properties of the observed data with hidden, strong assumptions, and the practioneer may not realize that the results of their data analysis reflect the assumptions they started with, simply because of complexity of the model. Second, such complex modeling almost always involves high-dimensional numerical computations using techniques like Markov Chain Monte Carlo, and it is not always evident how much of the resulting analysis depends on the “luck of the draw”. In principle, the latter aspect should not be present in any careful data analysis, but it is not always possible to verify software performance and functionality with big data, high-dimensional parameters and complex statistical models.

Hence, complex hierarchical models should be used with caution, and ought to be cross-checked using other models, robustness studies involving both prior and likelihood specifications, and multiple numerical approaches. Also, no matter how careful and diligent a researcher is, and how complex a model they might build, a parametric model is always limited by the imagination of the modeler, and can reflect only those properties of data that directly or indirectly captured by the assumed parametric process.

One potential alternative approach is to use non-parametric models, which are often infinite-dimensional parametric models. These have the attractive property of making less restrictive assumptions than finite-dimensional parametric models, hence they are better for capturing features of complex, high-dimensional phenomena, especially in large datasets. For example, instead of assuming that a conditional mean is linear in the covariates, we may assume it to be an arbitrary smooth function. Depending on the goal, it is sometimes possible to treat some data features as high-dimensional nuisance parameters that need not be explicitly modeled or estimated, for example, as was the case of Chatterjee *et al.* (2009). However, non-parametric models involve assumptions also, and some of the caveats mentioned above for parametric models are applicable in this case as well. For example, the assumption that a conditional mean is a smooth function may not be valid. Moreover, in essentially all situations, the estimation process and fitting of a non-parametric model is both less accurate, and less precise, than any applicable parametric alternative model. Thus, the property of robustness to parametric assumptions comes at a price of less accuracy and precision, which may defeat the purpose of a data analysis project.

An attractive middle path, compared to complex and obscure fully parametric modeling or inaccurate and imprecise non-parametric modeling, is something that mixes some components of both. While there are several existing semi-parametric modeling techniques available (Kosorok 2008), in this paper we propose a new approach, that is perhaps more suitable for analyzing complex, high-dimensional, and large sized data that routinely arise in climate studies. We propose to combine the framework of small area statistics with that of non-parametric curve fitting, to suggest a new semi-parametric approach. We discuss the details of the new approach in Section 2.3.

We discuss small area techniques and one standard non-parametric technique in Section 2. In Section 3, we present some simulation-based results and observations on our proposed scheme. The analysis for Arctic Ocean region seawater temperature is presented in Section 4. Then, we conclude the paper with Section 5.

In order to establish a context for our study, for the rest of this paper we consider temperature as the response variable, and restrict our attention to measurements taken at latitudes sixty degree North or higher, at depths up to 1500 meters, and on or after 1975. For convenience, we assign the name "the Arctic Ocean region" to the region of our study. We would like to emphasize that the data analysis technique developed for the climate study in this paper, and its associated statistical methodology, are applicable to any suitable response on the planet, and is not constrained by our choice of latitude, depth or time-frame. We illustrate our technique and present results for the Arctic Ocean region in this paper only for reasons of clarity, and because the Arctic Ocean region is one of the most climatologically sensitive areas on Earth.

2. NONPARAMETRIC REGRESSION USING SMALL AREA PERSPECTIVE

In this section, we first discuss in detail small area models. We then present our proposed approach of non-parametric curve estimation using small area methodology.

2.1 Small Area Models

The term small area may refer to geographical or spatial domains, demographic cohorts, or other divisions or groups of some natural data-indexing variable. The adjective small relates to the fact that the sample sizes corresponding to some or all of these groups may be limited. The art of small area statistical modeling lies in borrowing strength across the groups or small areas, so that the information from the limited sized direct sample in any given area is augmented by related information gleaned from other areas. This is traditionally done using a two-level hierarchical model. The second level linking model describes the model for carrying information between small areas, thus linking them together. Conditional on the second level, the first level (sampling model) describes the model for the limited sized direct sample (LSDS hereafter) for any given small area.

Perhaps the simplest, and most widely used small area framework is the Fay-Herriot (Fay and Herriot 1979) model, proposed to estimate the per-capita income of small places with population size less than 1000. This model is for univariate, area-level response Y_i corresponding to the i^{th} small area, for i =1, 2, ..., n. In addition to response Y_i, we also know the sampling model variance D_i, and covariates $\mathbf{x}_i \in \mathbb{R}^p$ for i = 1, 2, ..., n. Suppose that $(e_i, u_i) \overset{i.i.d.}{=} N_2(\mathbf{0}, \mathbf{I}_2)$ (the standard bivariate Normal distribution) for i = 1, 2, ..., n, and in terms of these random variables the Fay-Herriot model is given by

Level – I (sampling model)

$$[Y_i \,|\, \theta_i] = \theta_i + D_i^{1/2} e_i \sim N(\theta_i, D_i),$$

Level – II (linking model)

$$[\theta_i] = \mathbf{x}_i^T \beta + \psi^{1/2} u_i \sim N(\mathbf{x}_i^T \beta, \psi).$$

The unobserved random variables θ_1, ..., θ_n that denote the *true underlying effect* (TUE) all share the same marginal distribution, mean structure, and have a common variance ψ, and thus serve to relate all small areas together and help carry information between small areas. The prediction of the true underlying area effect (TUE), conditional on the data, $[(\theta_1, ..., \theta_n)|(Y_1, ..., Y_n)]$ is the target, and measuring the prediction error is an important related issue.

Much more detailed discussion on small area models in general, including usage, history, unit and area level models and so on may be found in Rao (2003) and Jiang and Lahiri (2006).

The above Fay-Herriot model makes some assumptions that need discussion. First, the small areas or groups are assumed to be unambiguous and clearly defined, and there is no overlap between the areas. This feature is quantified in the assumption that the random variable pairs (Y_i, θ_i) are independent for i =1, ..., n. However, in reality, geographical areas, demographic groups or other small areas may be expected to have some degree of dependence, overlapping groups are also a possibility. The notion of "what is a small area" may not be very well defined in some contexts.

Second, in the Fay-Herriot model we assume that the auxiliary variable **x** is linearly related to the response Y or the true underlying effect θ, the sampling or linking level variability do no depend on covariates, and both Level-I and Level-II random variables have the Normal distribution. These assumptions may not be all valid in practice.

2.2 Local Additive Polynomial Regression

Suppose $\{(Y_i, x_i) \in \mathcal{Y} \times \chi \subseteq \mathbb{R} \times \mathbb{R}^p, i = 1, 2, ..., n\}$is the data. Here, assume that χ is a compact subset of $\mathbb{R}^p$. Consider the model

$$Y(x) = \mu(x) + \sigma(x)e(x),$$

where $\{e(\cdot)\}$ is a mean zero, unit variance random field. Suppose our primary interest is in estimation of the mean function $\mu : \chi \to \mathcal{Y}$. The sample version of the above model is $Y_i = \mu(x_i) + \sigma(x_i)e_i$, where $\mathbb{E}e_i = 0, Ve_i = 1$. Generally, some more conditions are needed, for example, we may assume that $e_1, ..., e_n$ are independently and identically distributed as standard Normal random variables. Smoothness conditions on $\mu(\cdot)$ and $\sigma(\cdot)$ may be assumed also.

For the moment assume $p = 1$, *i.e.*, there is only one covariate. Corresponding to the i^{th} observation (Y_i, x_i), for each fixed $x \in \mathcal{X} \subset \mathbb{R}$ and each non-negative integer J, define

$$(\beta(x, J) \equiv)\beta = (\beta_0, ..., \beta_J) \in \mathbb{R}^{J+1}$$

$$v_i(x, \beta) = \sum_{j=0}^{J} \beta_j \frac{(x_i - x)^j}{j!}$$

Consider the problem of minimizing the following function:

$$\psi_{x,J,h}(\beta) = \sum_{i=1}^{n} (Y_i - v_i(x, \beta))^2 K_h(x_i - x),$$

where $K_h(t) \equiv K(t/h)$ is smoothing kernel function. For example, the Gaussian kernel is given by $K_h(t) = \exp\{-t^2/(2h^2)\}$.

The value of β, say $\hat{\beta} \equiv \hat{\beta}(x; J, h)$ at which the minimization of $\psi_{x,J,h}$ takes place is naturally dependent on the choices of $x \in \chi$, J, and the smoothing parameter h. Note that $\hat{\beta}_0(x; J, h)$ is a local polynomial estimate of $\mu(x)$, the expectation of the response corresponding to covariate value x (Fan and Gijbels 1996).The quality of this estimate depends on the choice of the tuning parameters J and h, and to some extent, the nature of the kernel function $K(\cdot)$.

The above minimization is motivated by considering the fact $\mathbb{E}Y_i = \mu(x_i)$ may be assumed to allow a Taylor series expansion at x:

$$\mathbb{E}Y_i = \mu(x_i) = \sum_{j=0}^{J} \frac{d^j \mu(x)}{(dx)^j} \frac{(x_i - x)^j}{j!} + R(x, J)$$

$$= \sum_{j=0}^{J} \beta_j \frac{(x_i - x)^j}{j!} + R(x, J).$$

Here, $R_{x,\,J}$ is the remainder term, satisfying

$$R(x, J) = \frac{d^{J+1}\mu(x + d(x_i - x))}{(dx)^{J+1}} \frac{(x_i - x)^{J+1}}{(J+1)!}$$

for some $d \in (0, 1)$. This shows that $\hat{\beta}_j(x; J, h)$ is an estimate of the j^{th} derivative of $\mu(x)$, *i.e.* $\mu_j(x) = \frac{d^j \mu(x)}{(dx)^j}$. Let us denote the above scheme as model $\mathcal{M}(J, h)$, in recognition that the results of this procedure depends on J and h. Note that if our model assumptions allow, an estimate of $R_{x', J}$ in model $\mathcal{M}(J, h)$ might be obtained by simply fitting model $\mathcal{M}(J + 1, h)$ and obtaining $R_{x,\,J} \approx \hat{\beta}_{J+1} \frac{(x_i - x)^{J+1}}{(J+1)!}$. In other words, for every J, the last term in $v_i(x, \beta)$ is close to the approximation error for model $\mathcal{M}(J, h)$.

For the case of p-dimensional covariates, we may proceed along similar lines to those outlined above, using a multi-variate Taylor series expansion. A dimension-reduced alternative, in case we have p covariates $x_{\square} = (x_1, ..., x_p)$, is to use a generalized additive model (Hastie and Tibshirani 1990), where we have a further assumption that

$$\mu(x_{\square}) = \sum_{k=1}^{p} \tilde{\mu}_k(x_k),$$

where $\tilde{\mu}_k(\cdot)$ is a function of the k^{th} covariate only.

We now describe how to use the generalized additive model in a local polynomial framework. The basic concept behind this model has been mentioned earlier, see Hastie *et al.* (2009), Chapter 9, for example. Suppose x_{ik} is the i^{th} observation for the k^{th} covariate, tagged to the i^{th} response Y_i.

For each fixed non-negative integer J, smoothing parameter $h > 0$, and each $x_{\square} = (x_1, ..., x_p) \in \mathcal{X} \subset \mathbb{R}^p$, we define

$$(\beta(x_{\square}, J) \equiv)\beta = (\beta_0, \beta_{11}, ..., \beta_{1p}, \beta_{21}, ..., \beta_{2p}, ..., \beta_{J1}, ..., \beta_{Jp}) \in \mathbb{R}^{Jp+1}$$

$$v_i(x_{\square}, \beta) = \sum_{j=0}^{J} \sum_{k=1}^{p} \beta_{jk} \frac{(x_{ik} - x_k)^j}{j!}$$

and minimize

$$\psi_{x_{\square}j,h}(\beta) \;=\; \sum_{i=1}^{n}(Y_i - v_i(x_{\square},\beta))^2 K_h(x_{i\square} - x_{\square}),$$

where here K_h is a p-dimensional kernel. We call the above scheme the *local additive* polynomial regression (LAPR) technique. This is the special case of multivariate local polynomial regression where the cross-derivatives of $\mu(\cdot)$ are assumed to be zero, as well as a case of a generalized additive model fitting using local polynomial method. We will use LAPR as our nonparametric regression fitting methodology for the rest of this paper.

As with any nonparametric regression, the tuning parameters (J and h in this case) are important in determining the quality of the fit. Also note that $v_i(x_{i\square}, \beta)$ is the estimate for $\mu(x_{i\square})$, a fact that we will use in the sequel.

2.3 The Proposed Model: SmAr-LAPR

Note that the above scheme of local additive polynomial regression, or LAPR, purports to find $\hat{\mu}(x)$ for each $x \equiv x_{\square} \in \mathcal{X} \subset \mathbb{R}^p$ through the estimated intercept term $\hat{\beta}_0(x)$ in each model $\mathcal{M}(J, h)$. But in reality, model $\mathcal{M}(J, h)$ essentially substitues $v_i(x_{\square}, \beta)$ $= \sum_{j=0}^{J}\sum_{k=1}^{p}\beta_{jk}\frac{(x_{ik}-x_k)^j}{j!}$ in place of $\mu(x_i)$ in the expression of $\psi_{x,J,h}(\beta)$, thus effectively ignoring that the Taylor series remainder term $R(x, J)$. This is an essential step in making this a well-defined problem. However, a careful study of this problem suggests that this process has two sources of error: the noise terms e_i arising out of the random field $e(\cdot)$, which is an usual component in statistics, and the approximation error $R(x, J)$. Let us use the notation $v(x_{\square})$ as a generic term for the finite-dimensional parameter based approximation of $\mu(x_{\square})$.

Based on the above reasoning, we might propose the theoretical model corresponding to LAPR as a Fay-Herriot type hierarchical model as follows: for every model $\mathcal{M}(J, h)$, we have:

Level – I (sampling model):

$$[Y(x_{\square})|\mu(x_{\square}),\, \sigma(x_{\square})] = \mu(x_{\square}) + \sigma(x_{\square})e_{x_{\square}}$$
$$\sim N(\mu(x_{\square}),\, \sigma^2(x_{\square})),$$

([Observation | True infinite dimensional parameters] ~ N (True mean, True var).)

Level – II (approximation model):

$$[\mu(x_{\square})] = v(x_{\square}) + \psi^{1/2}(x_{\square})u_i \sim N(v(x_{\square}),\, \psi(x_{\square})),$$

([True mean] ~ N (Estimable approximate mean, Approximation error var).)

In this Fay-Herriot type nonparametric regression model, the first layer, Level-I, characterizes the nature of the responses $Y(x_{\square})$ in terms of the infinite dimensional (and hence not completely estimable from finite sample data) unknown parameters, the mean function $\mu(x_{\square})$ and the standard deviation function $\sigma(x_{\square})$. The second layer characterizes the workable version of nonparametric regression, where $\mu(x_{\square})$ is approximated by a workable, finite-parameter approximate function, with an allowance for approximation error characterized by the Level-II randomness and variance ψ. This is a multi-index model as $x_{i\square}$ is a multi-dimensional covariate.

The above generic model holds for all viable choices of J and h, and this allows for model-averaged estimates for $v(x_{\square})$ and ψ. As discussed earlier, we estimate $v_i(x_{\square}, \beta)$ in model $\mathcal{M}(J, h)$ by minimizing

$$\psi_{x_{\square},J,h}(\beta) = \sum_{i=1}^{n}(Y_i - v_i(x_{\square},\beta))^2 K_h(x_{i\square} - x_{\square}),$$

Let us define the model averaged Level-II mean function as

$$\hat{v}(x_{\square}) = |\mathcal{M}(J,h)|^{-1} \sum_{\mathcal{M}(J,h)} \hat{v}_i(x_{\square},\beta)$$

This is simply the average of the Level-II mean from all the models under consideration.

We might have considered more complex schemes for averaging across models, but opted for simplicity at this stage.

Similarly, we define the model averaged Level-II variance function as

$$\psi(x_{\square}) = |\mathcal{M}(J,h)|^{-1} \sum_{\mathcal{M}(J,h)} (\hat{v}_i(x_{\square},\beta) - \hat{v}(x_{\square}))^2.$$

This variance describes the model to model variability at each $x \in \mathcal{X}$.

The residual for the i^{th} observation is given by

$$r_i = Y_i - \hat{v}_i(x_{\square}, \boldsymbol{\beta}),$$

and notice that $\mathbb{E} r_i^2 \approx \psi(x_{\square}) + \sigma^2(x_{\square})$. This relation allows us to propose that

$$\sigma^2(x_{\square}) = r_i^2 - \psi.$$

Note that the above techniques of using least squares for estimating the mean, and for decomposing the observed variability into Level-I and Level-II variability, is similar to the Prasad-Rao method (Prasad and Rao 1990) for the Fay-Herriot model.

The above framework now allows for prediction of the mean function $\mu(x_{\square})$ at $x_{\square}$, $i = 1, ..., n$: using

$$\hat{\mu}(x_{\square}) = [\sigma^2(x_{\square}) + \psi^{-1}(x_{\square})]^{-1} (\sigma^{-2}(x_{\square})Y(x_{\square}) + \psi^{-1}(x_{\square})\, \hat{v}(x_{\square})). \qquad (2.1)$$

Here however, the availability of $Y(x_{\square})$ depends on whether $(x_{\square})$ has been observed in the data or not, so some practical alternatives are needed. We suggest some alternatives below:

$$\hat{\mu}_1(x_{\square}) = \begin{cases} [\sigma^{-2}(x_{\square}) + \psi^{-1}(x_{\square})]^{-1} \\ (\sigma^{-2}(x_{\square})Y(x_{\square}) + \psi^{-1}(x_{\square})\hat{v}(x_{\square})), \\ \qquad x_{\square} \in \{x_{1\square}, ..., x_{n_{Box}}\}, \\ [\sigma^{-2}(x_{\square}) + \psi^{-1}(x_{\square})]^{-1} \\ (\sigma^{-2}(x_{\square})\hat{v}(x_{\square}) + \psi^{-1}(x_{\square})\hat{v}(x_{\square})), \\ \qquad x_{\square} \not\subset \{x_{1\square}, ..., x_{n_{Box}}\}, \end{cases} \qquad (2.2)$$

$$\hat{\mu}_{2Jh}(x_{\square}) = \begin{cases} [\sigma^{-2}(x_{\square}) + \psi^{-1}(x_{\square})]^{-1} \\ (\sigma^{-2}(x_{\square})Y(x_{\square}) + \psi^{-1}(x_{\square})\hat{v}(x_{\square})), \\ \qquad x_{\square} \in \{x_{1\square}, ..., x_{n_{Box}}\}, \\ [\sigma^{-2}(x_{\square}) + \psi^{-1}(x_{\square})]^{-1} \\ (\sigma^{-2}(x_{\square})\hat{v}_i(x_{\square}) + \psi^{-1}(x_{\square})\hat{v}(x_{\square})), \\ \qquad x_{\square} \not\subset \{x_{1\square}, ..., x_{n_{Box}}\}, \end{cases} \qquad (2.3)$$

$$(2.4)$$

$$\hat{\mu}_{3Jh}(x_{\square}) = [\sigma^{-2}(x_{\square}) + \psi^{-1}(x_{\square})]^{-1}$$

$$(\sigma^{-2}(x_{\square})\hat{v}_i(x_{\square}) + \psi^{-1}(x_{\square})\hat{v}(x_{\square})), \forall x \in \mathcal{X}. \qquad (2.5)$$

A brief discussion is in order on these three predictors. First, (2.2) and (2.3) are a convex combination of the original observations and a smoothed predictor. This can be thought of as a blend of nearest neighbor and kernel smoothing methods. These are obviously only available for those $x_{i_{\square}}$ which have been observed, since these require Y_i. The predictors $\hat{\mu}_1(x_{\square})$ and $\hat{\mu}_{2Jh}(x_{\square})$ differ on how unobserved $(x_{\square})$ is handled. In the former, the model-averaged mean is used, while in the latter case, the fit from model $\mathcal{M}(J, h)$ is used, consequently it varies from model to model. The third kind of predictor, $\hat{\mu}_{3Jh}(x_{\square})$ is a convex combination of the fit from model $\mathcal{M}(J, h)$ and the model-averaged prediction.

Motivating from the Fay-Herriot model, in the current framework each model $\mathcal{M}(J, h)$ serves as an "area". The TUE is the true unknown mean function, which does not depend on the area. In particular, we broaden the concept of an "area", and that of a "true underlying effect" (TUE) that are central to small area modeling.

Notice that the basic predictor $\hat{\mu}(x_{\square})$ can be thought of as a conditional expectation.

This predictor is associated with the conditional variance

$$[\sigma^{-2}(x_{\square}) + \psi^{-1}]^{-1} < \min\{\sigma^2(x_{\square}), \psi\}.$$

Herein lies the advantage of using a small area model for prediction: the predictive variance is lower than the minimum of both Level-I and Level-II variances.

3. SOME SIMULATION-BASED EXAMPLES

We consider six models in all: we used $J = 0$ and $J = 1$, and two bandwidths $h = 0.5$ and another h obtained by the default method due to Silverman (1998). In addition, we used linear and quadratic regression as two fully parametric models, to serve as baselines. For each model, we have the model-specific prediction, and the three blended predictions given in (2.2), (2.3) and (2.5). Hence, there are twenty-four predictors in all.

We report the performance of each of the twenty-four predictors in three parts: as mean squared errors for the entire data, for in-sample mean squared errors, and out-of-sample mean squared errors.

We use a four-dimensional covariate for the simulations. We consider the model:

Table 5.1 Mean squared errors (MSE) from 24 predictors: total, in-sample and out-of-sample values.

Predictor	Total	In-Sample	Out-Sample
Deg = 0, h = 0.5	5431.39	5571.06	4872.68
Deg = 1, h = 0.5	5457.94	5594.89	4910.14
Deg = 0, h = Sl	8866.82	8963.11	8481.67
Deg = 1, h = Sl	5386.95	5539.18	4778.02
Lin	2508.78	2565.86	2280.48
Quad	1732.42	1699.90	1862.51
Deg = 0, h = 0.5 (μ_1)	1570.23	1136.79	3303.99
Deg = 1, h = 0.5 (μ_1)	1573.19	1140.49	3303.99
Deg = 0, h = Sl (μ_1)	1886.30	1531.88	3303.99
Deg = 1, h = Sl (μ_1)	1565.26	1130.58	3303.99
Lin (μ_1)	1146.20	606.75	3303.99
Quad (μ_1)	995.51	418.40	3303.99
Deg = 0, h = 0.5 (μ_2)	2329.73	1693.99	4872.68
Deg = 1, h = 0.5 (μ_2)	2387.08	1756.31	4910.14
Deg = 0, h = Sl (μ_2)	4777.48	3851.43	8481.67
Deg = 1, h = Sl (μ2)	2372.14	1770.67	4778.02
Lin (μ_2)	785.39	411.62	2280.48
Quad (μ_2)	593.05	275.68	1862.51
Deg = 0, h = 0.5 (μ_3)	4425.63	4500.77	4125.06
Deg = 1, h = 0.5 (μ_3)	4492.06	4567.08	4191.98
Deg = 0, h = Sl (μ_3)	6404.90	6394.04	6448.34
Deg = 1, h = Sl (μ_3)	4490.20	4569.35	4173.62
Lin (μ_3)	3180.44	3282.67	2771.51
Quad (μ_3)	3306.09	3411.56	2884.23

$$\begin{aligned} Y_i = {} & \exp(1.2X[i, 1]) + 10 \sin(1.05X[i, 3]) \\ & + 2X[i, 1] + X[i, 2] + 0.5X[i, 3] + 0.25X[i, 4] \\ & + e[i],\ i = 1, ..., n \end{aligned}$$

with a sample size of $n = 500$. We randomly select 20% of data as hold-out for out-of-sample evaluation. Table 5.1 provides the results for this example. Notice that, as a whole, blended predictors like $\hat{\mu}_1(\cdot)$ do very well across the board. Predictors like linear or quadratic regression perform well also when blended into $\hat{\mu}_1(\cdot)$ or $\hat{\mu}_2(\cdot)$.

We present some plots for comparison. In Fig. 5.1 we present the plots for the six versions of $\hat{\mu}_1(\cdot)$ (corresponding to the six models). Each plot in that figure has covariate $X[, i]$, $i = 1, 2, 3, 4$ in the x-axis. Owing to similarity in predictions from some of the models.

In Fig. 5.2 we compare the "degree = 1, $h = 0.5$" fits for the four kinds of models. The different curves represent the unblended local polynomial fit, and $\hat{\mu}_j(\cdot)$, $j = 1, 2, 3$. Each plot in that figure has covariate $X[, i]$, $i = 1, 2, 3, 4$ in the x-axis. Notice that the usual local linear regression fit is a poor fit in this case. On the other hand, $\hat{\mu}_3(\cdot)$ seems to overfit. The $\hat{\mu}_1(\cdot)$ or $\hat{\mu}_2(\cdot)$ curves seem to do better both in capturing the shape of the curve as well as not overfitting.

In Fig. 5.3 we compare the linear regression fits for the four kinds of models. The different curves represent the unblended linear regression fit, and $\hat{\mu}_j(\cdot)$, $j = 1, 2, 3$. Each plot in that figure has covariate $X[, i]$, $i = 1, 2, 3, 4$ in the x-axis. Our observations are similar to that of the previous figure.

More simulations were carried out, that are not reported here. We used several alternative mean and variance functions, and did replicated experiments to study the effect of chance factors. Various choices of degree of fitted curve and bandwidths were used. Our general observations are similar to the descriptions above: that $\hat{\mu}_1(\cdot)$, and many times $\hat{\mu}_2(\cdot)$ curves seem to capture the shape of the curve as well as prevent against overfitting. Both these curves have an interesting explanation: they represent a weighted combination of a nearest neighbor curve fit and a local polynomial curve fit. All the blended curves are not particularly smooth, a fact illustrated in the spikes on the presented figures. This suggests our blended Fay-Herriot type curves can possibly capture non-smooth shapes as well. Very interestingly, the blended curves $\hat{\mu}_j(\cdot)$, $j = 1, 2, 3$ also seem to be remarkably robust to bandwidth selection, and to some extent to degree specification.

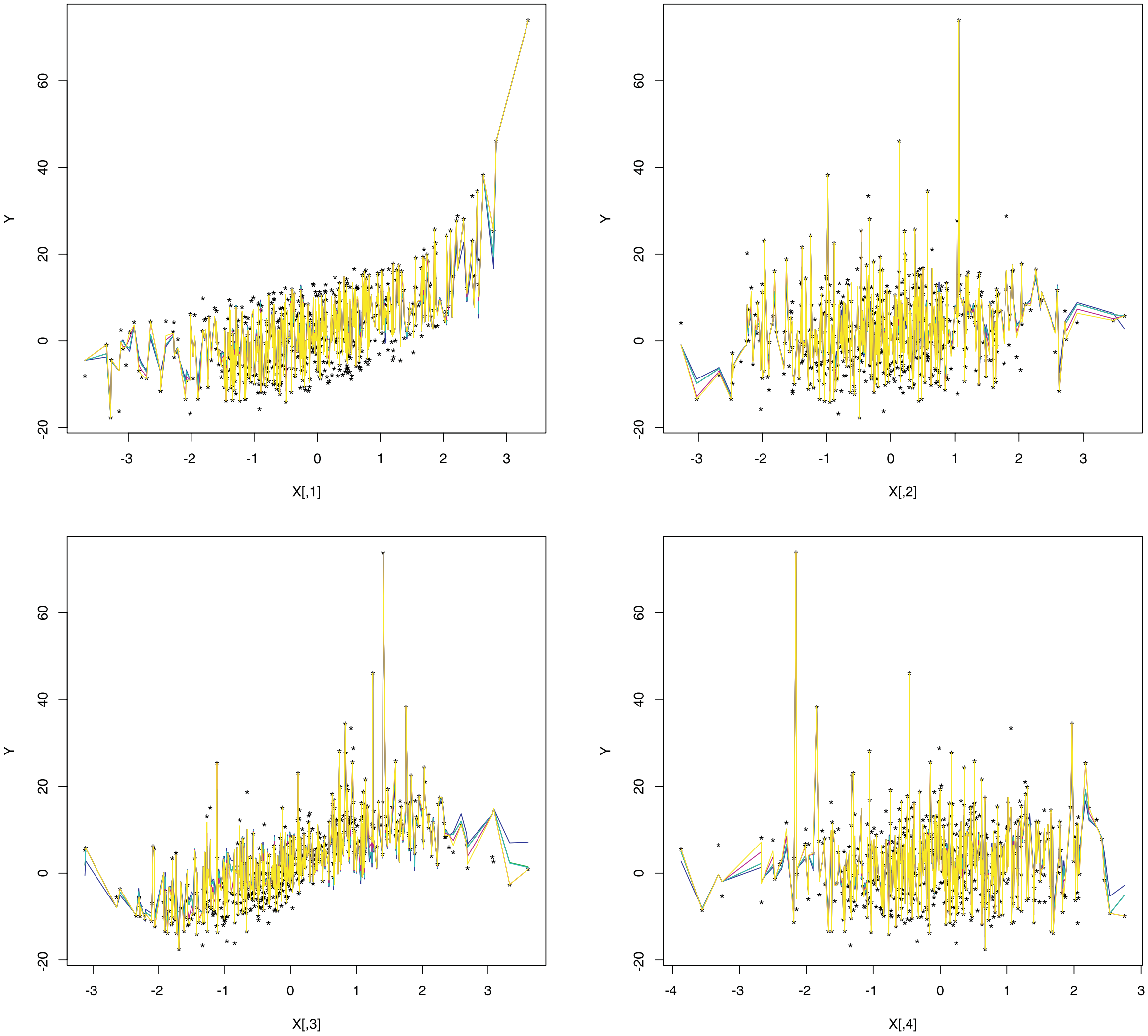


Fig. 5.1. Simulated data: curves of various $\mu_1(\cdot)$ predictors. (Red:degree = 0, h = 0.5, green:degree = 1, h = 0.5, blue:degree = 0, h = Silverman, cyan:degree = 0, h = Silverman, magenta:linear, yellow:quadratic)

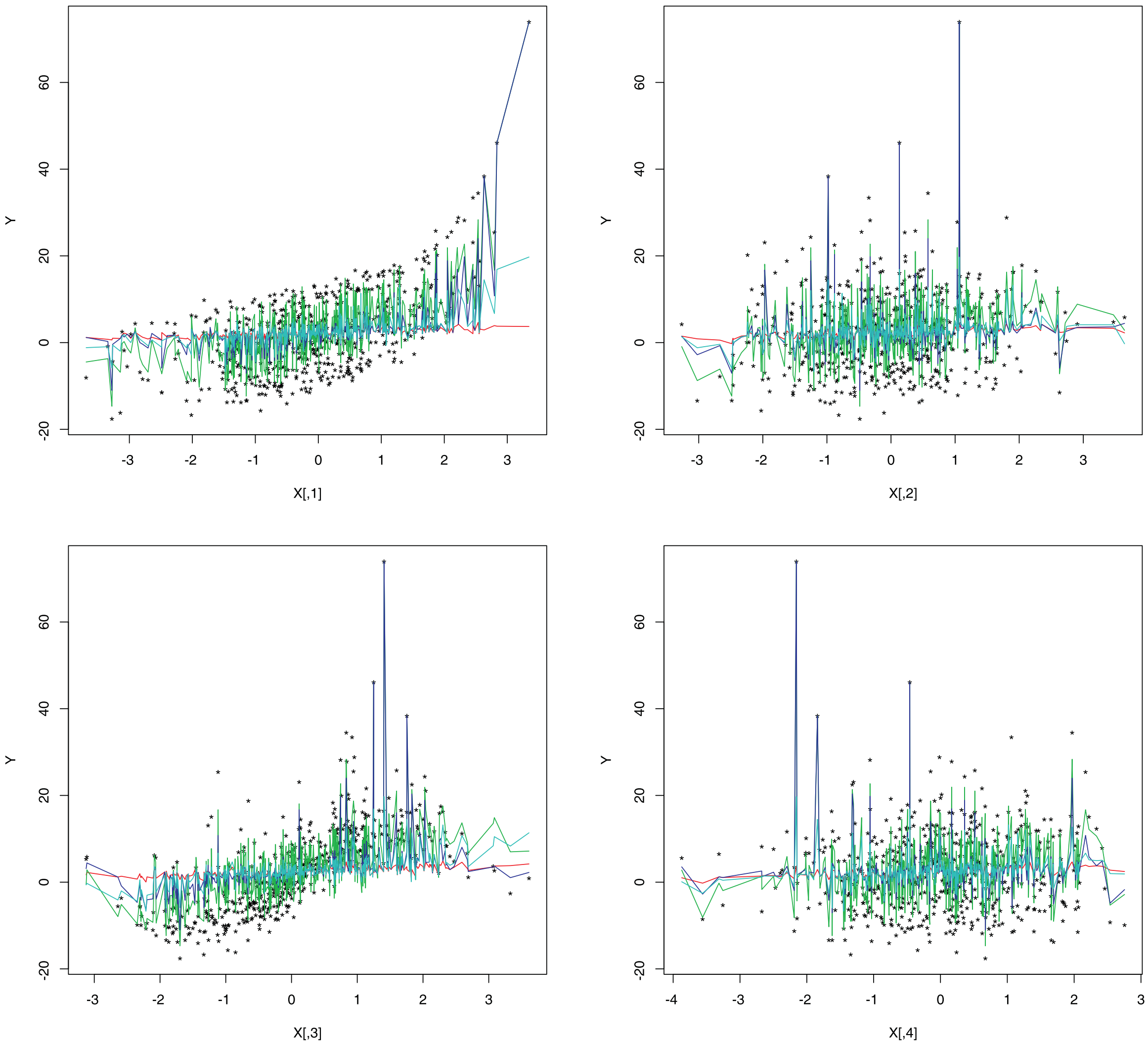


Fig. 5.2. Simulated data: curves of various kinds of predictors. We use the degree = 1, h = 0.5 case. (Red:local polynomial, green: $\hat{\mu}_1(\cdot)$, blue: $\hat{\mu}_2(\cdot)$, cyan: $\hat{\mu}_3(\cdot)$)

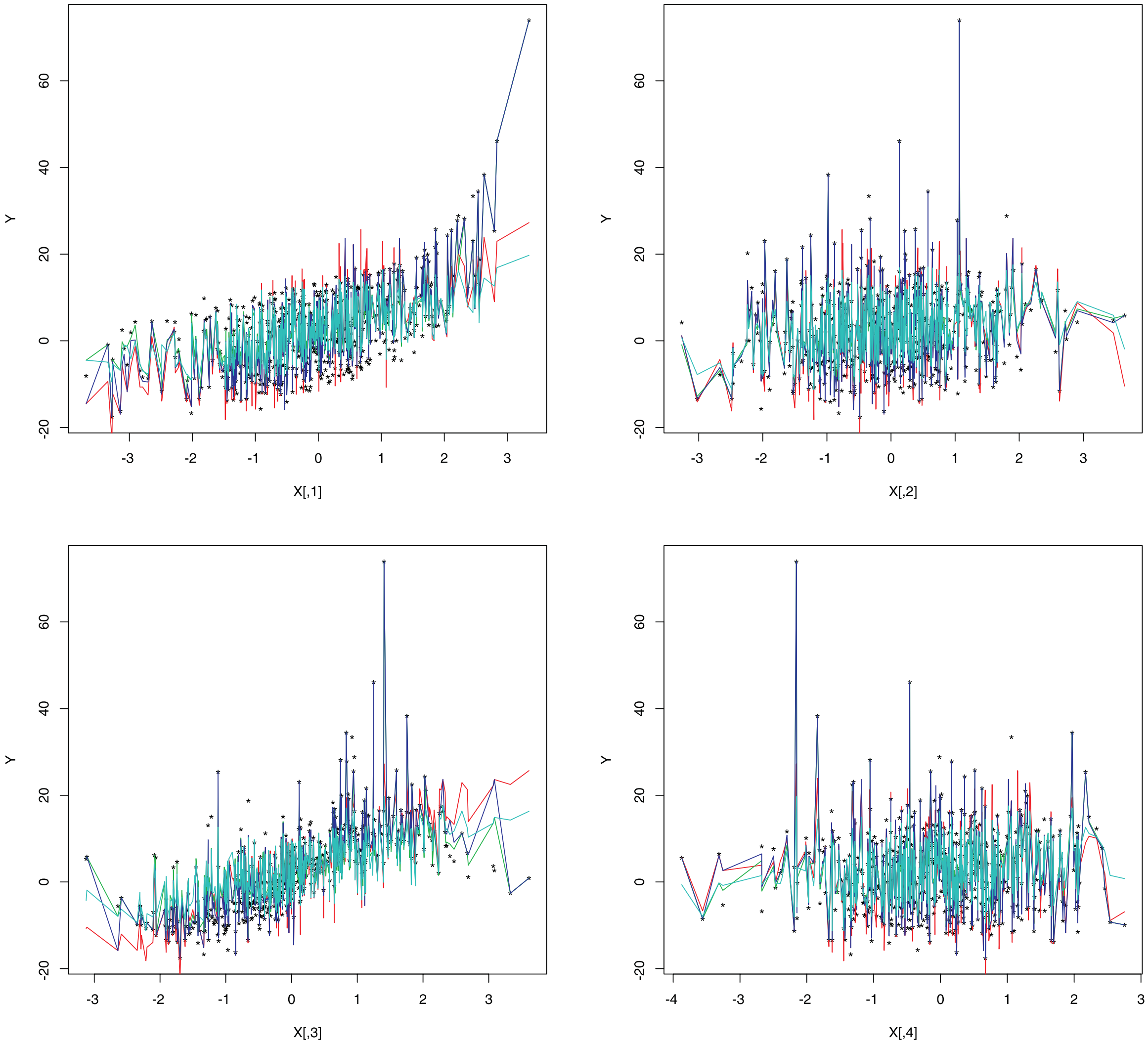


Fig. 5.3. Simulated data: curves of various kinds of predictors. We use the linear regression case. (Red: regression, green: $\hat{\mu}_1(\cdot)$, blue: $\hat{\mu}_2(\cdot)$, cyan: $\hat{\mu}_3(\cdot)$)

Table 5.2. Mean squared errors (MSE) from 24 predictors: total, in-sample and out-of-sample values in the temperature of sea water in the Arctic Ocean region.

Predictor	Total	In-Sample	Out-Sample
Deg = 0, h = 0.5	120.63	116.41	137.47
Deg = 1, h = 0.5	95.89	94.51	101.36
Deg = 0, h = Sl	131.22	125.35	154.66
Deg = 1, h = Sl	221.77	228.28	195.80
Lin	44.26	45.30	40.12
Quad	25.29	25.06	26.19
Deg = 0, h = 0.5 (μ_1)	37.21	27.13	77.48
Deg = 1, h = 0.5 (μ_1)	33.43	22.40	77.48
Deg = 0, h = Sl (μ_1)	38.77	29.07	77.48
Deg = 1, h = Sl (μ_1)	48.35	41.05	77.48
Lin (μ_1)	25.18	12.08	77.48
Quad (μ_1)	20.08	5.70	77.48
Deg = 0, h = 0.5 (μ_2)	57.51	37.49	137.47
Deg = 1, h = 0.5 (μ_2)	53.54	41.57	101.36
Deg = 0, h = Sl (μ_2)	67.27	45.38	154.66
Deg = 1, h = Sl (μ_2)	111.59	90.49	195.80
Lin (μ_2)	18.86	13.53	40.12
Quad (μ_2)	8.19	3.68	26.19
Deg = 0, h = 0.5 (μ_3)	89.69	87.01	100.38
Deg = 1, h = 0.5 (μ_3)	90.33	88.95	95.82
Deg = 0, h = Sl (μ_3)	97.26	93.69	111.52
Deg = 1, h = Sl (μ_3)	133.94	132.42	140.01
Lin (μ_3)	71.26	69.95	76.49
Quad (μ_3)	64.41	62.80	70.87

4. ARCTIC OCEAN SEAWATER TEMPERATURE DATA ANALYSIS

We apply the techniques described above on the Arctic Ocean region seawater temperature data. We use Time, as defined by a Year-Month combination, logarithm of depth, and trigonometric transformations of latitude and longitude as covariates. We standardize the covariates for better fitting. Similar to the simulation example, we hold out a randomly selected 20% of data for out-of-sample prediction evaluation. The results from the mean squared error analysis is presented for all the twenty-four models in Table 5.2. Plots comparable to the simulation section are presented in Fig. 5.4, Fig. 5.5 and Fig. 5.6.

Note that $\hat{\mu}_2(\cdot)$ with quadratic and linear fits, and $\hat{\mu}_1(\cdot)$ with the quadratic fit seem very good fits and considerably better than other models. Hence, we consider these, along with $\hat{\mu}_1(\cdot)$ with the linear fit as another option, for an observed-versus-predicted plot. We divide this figure into two parts, one for those data points that were in the sample, and another for those points that were not used in the model construction. This result is presented in Fig. 5.7. It can be seen that $\hat{\mu}_2(\cdot)$ with quadratic fit, which appears to be the best predictor from Table 5.2, does not capture the patterns particularly well in either in-sample or out-of-sample data. The predictor $\hat{\mu}_2(\cdot)$ with linear fit also performs poorly in out-of-sample pattern matching. On the other hand, predictors $\hat{\mu}_1(\cdot)$ with linear or quadratic fit seem to match patterns much better. However, their poorer numeric performance is a reflection of having less accuracy and precision on an average.

The better prediction based on linear and quadratic fits suggest that there are more structures in the data than what we have evinced so far. This is also in keeping with what Bhattacharjee and Chatterjee (2013) obtained from a Bayesian perspective.

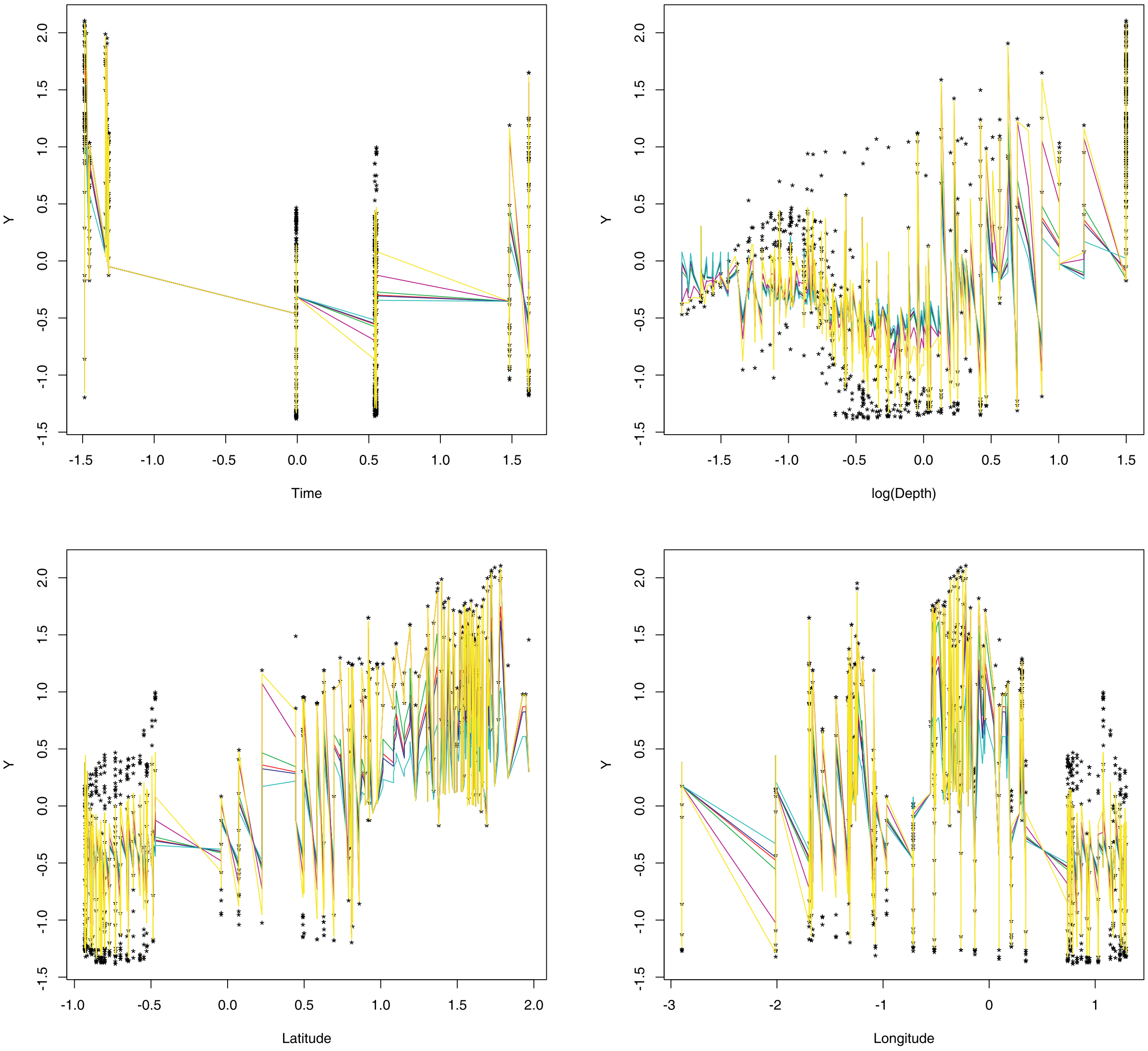


Fig. 5.4. Arctic Ocean region water temperature data: curves of various $\mu_1(\cdot)$ predictors. (Red:degree = 0, *h* = 0.5, green:degree = 1, h = 0.5, blue:degree = 0, *h* = Silverman, cyan:degree = 0, *h* = Silverman, magenta:linear, yellow:quadratic)

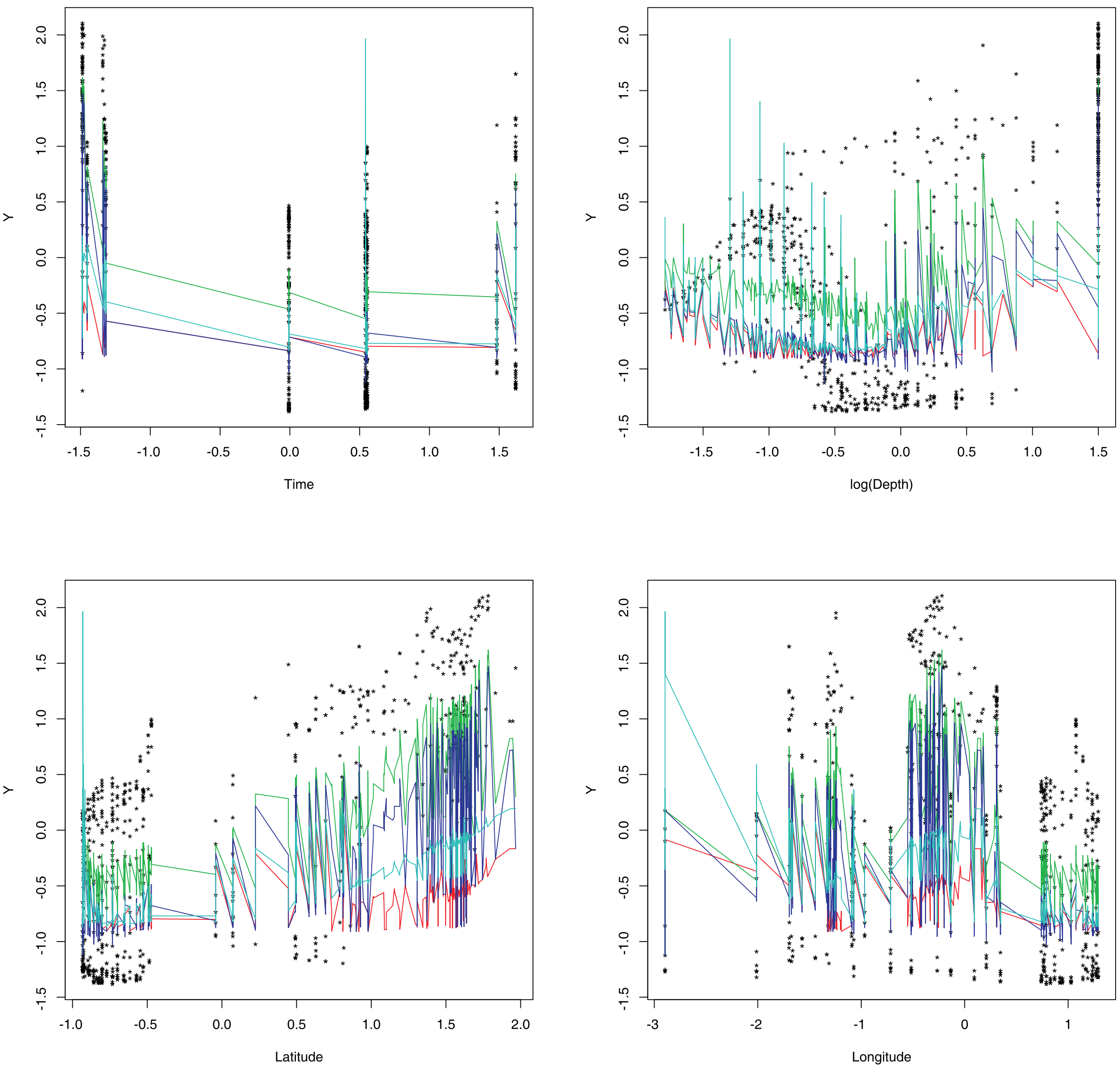


Fig. 5.5. Arctic Ocean region water temperature data: curves of various kinds of predictors. We use the degree = 1, h = 0.5 case. (Red:local polynomial, green: $\hat{\mu}_1(\cdot)$, blue: $\hat{\mu}_2(\cdot)$, cyan: $\hat{\mu}_3(\cdot)$)

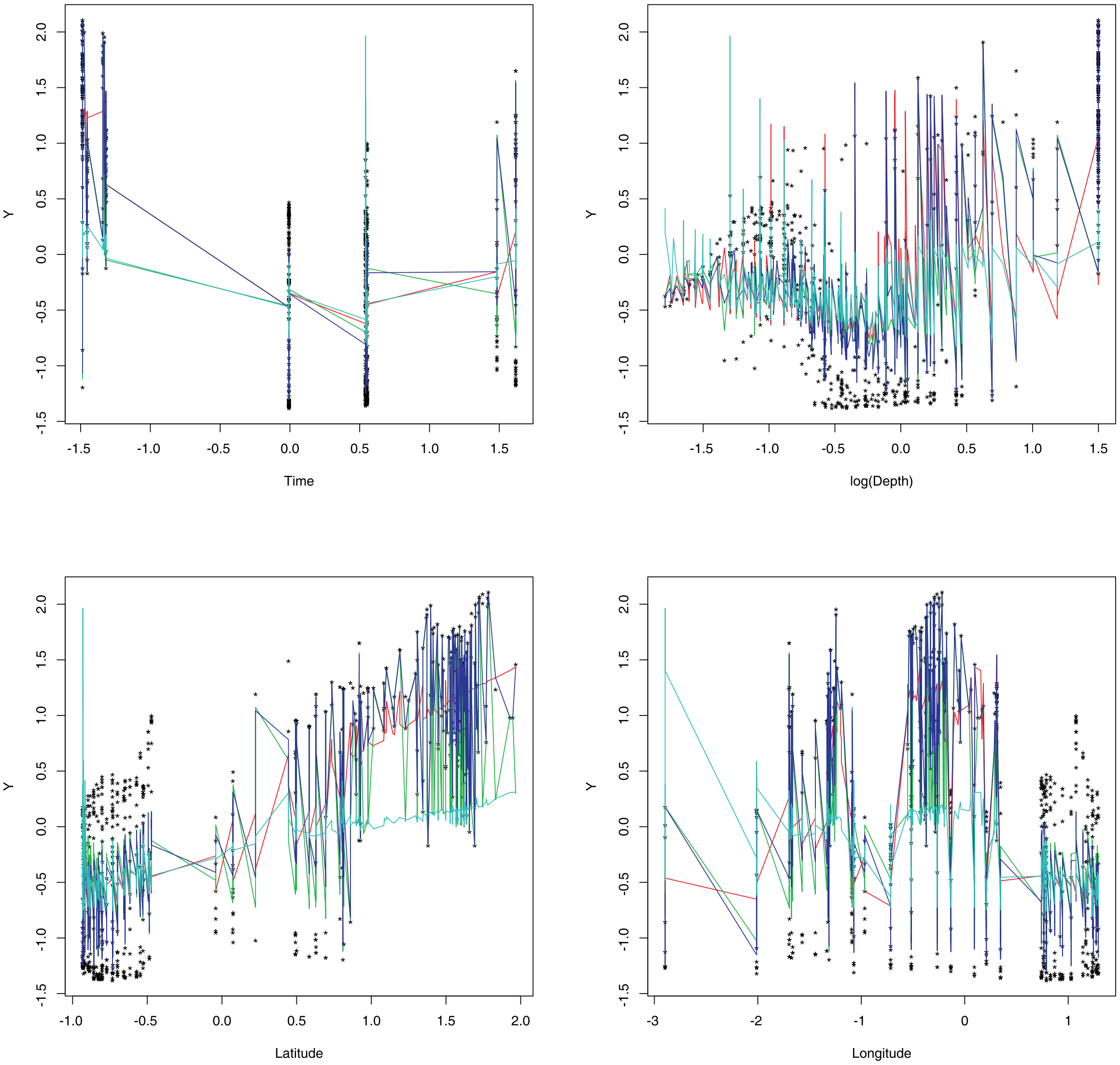
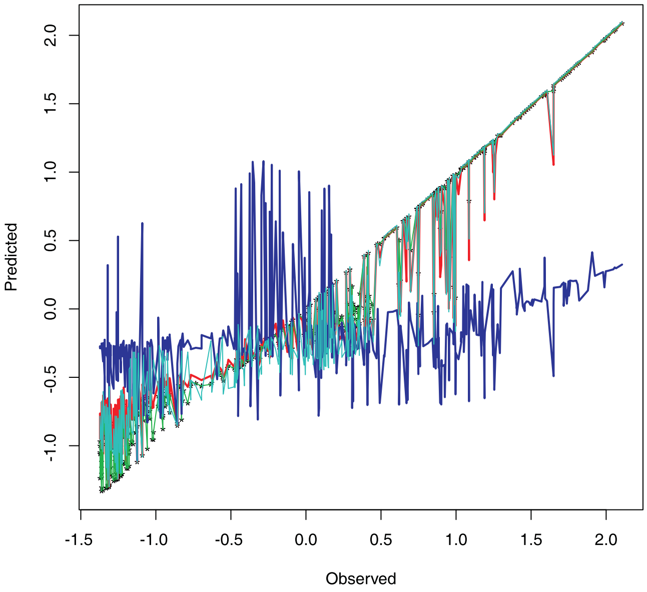


Fig. 5.6. Arctic Ocean region water temperature data: curves of various kinds of predictors. We use the linear regression case. (Red: regression, green: $\hat{\mu}_1(\cdot)$, blue: $\hat{\mu}_2(\cdot)$, cyan: $\hat{\mu}_3(\cdot)$)



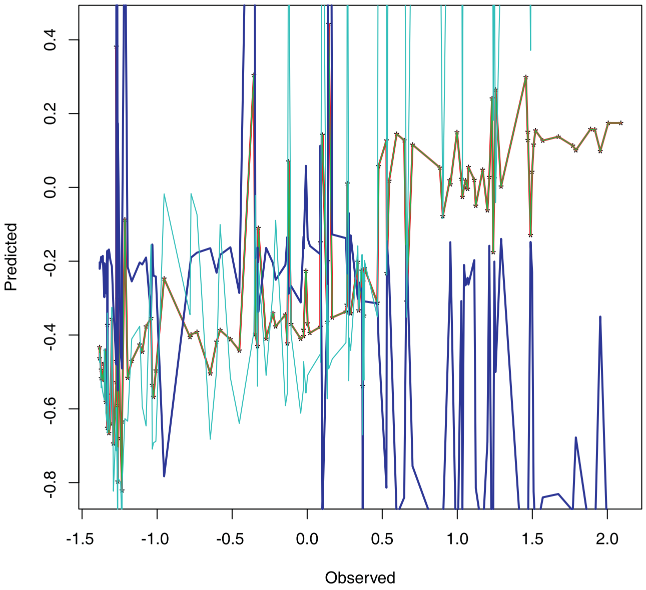


Fig. 5.7. Arctic Ocean region water temperature data: curves of observed versus predicted data. Top figure is in-sample prediction, bottom figure is out-of-sample prediction. (Red: $\hat{\mu}_2(\cdot)$ with linear regression blend, green: $\hat{\mu}_1(\cdot)$ with quadratic regression blend, blue: $\hat{\mu}_2(\cdot)$ with linear regression blend, cyan: $\hat{\mu}_2(\cdot)$ with quadratic regression blend.)

5. CONCLUSION

We have proposed a way of blending in nonparametric regression with the Fay-Herriot type small area model. This can be used in two ways at least: to improve regression performance, or to use non-parametric surve fitting in small area problems. In this paper, we have illustrated the former. We have proposed three kinds of blended curves. In our simulation examples, $\hat{\mu}_1(\cdot)$, and many times $\hat{\mu}_2(\cdot)$ curves seem to capture the shape of the curve as well as prevent against overfitting, while $\hat{\mu}_3(\cdot)$ seems to do quite well many times. The curves $\hat{\mu}_1(\cdot)$ and $\hat{\mu}_2(\cdot)$ represent a weighted combination of a nearest neighbor curve fit and a local polynomial curve fit, which might explain their good performance. All the blended curves are not particularly smooth, which suggests our blended Fay-Herriot type curves can possibly capture non-smooth shapes as well. The blended curves $\hat{\mu}_j(\cdot)$, j = 1, 2, 3 also seem to be robust to bandwidth selection, and to some extent to degree specification.

ACKNOWLEDGEMENTS

The second authors research was partially funded by NSF grants # SES-0851705 and # IIS-1029711, and research grants from the Institute on the Environment and College of Liberal Arts, University of Minnesota.

REFERENCES

Allen, M.R. and Tett, S.F.B. (1999). Checking for model consistency in optimal fingerprinting. *Clim. Dyn.*, **15**, 419-434.

Bhattacharjee, M. and Chatterjee, S. (2013). On Bayesian spatio-temporal modeling of oceanographic climate characteristics. *preprint*.

Chatterjee, S., Deng, Q., and Xu, J. (2009). The statistical evidence of climate change: an analysis of global seawater data. Technical Report #677, School of Statistics, University of Minnesota.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman and Hall, London, 341.

Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedure to census data. *J. Amer. Statist. Assoc.*, **74,** 269-277.

Hasler, Natalia, Werth, David and Avissar, Roni (2009). Effects of tropical de-forestation on global hydroclimate: A multimodel ensemble analysis. *J. Climate*, **22**, 1124-1141.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman and Hall, CRC.

Hastie, T.R., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* (Second Edition), Springer.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation (with discussions). *Test*, **15(1)** 1-96.

Kosorok, M.R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.

Lee, Terry, C.K., Francis, W. Zwiers, Gabriele, C. Hegerl, Xuebin Zhang and Min, Tsao (2005). A bayesian climate change detection and attribution assessment. *J. Climate*, **18,** 2429-2440.

Min, Seung-Ki and Andreas, Hense. (2006). A bayesian assessment of climate change using multimodel ensembles. Part I: Global Mean Surface Temperature. *J. Climate,* **19,** 3237-3256.

Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *J. Amer. Statist. Assoc*., **85,** 163-171.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley, New York.

Schmidt, G.A., Bigg, G.R. and Rohling, E.J. (1999). Global Seawater Oxygen-18 Database. http://data.giss.nasa.gov/o18data/

Silverman, B.W. (1998). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, CRC, London, 48.