

# Gaussian Semiparametric Analysis Using Hierarchical Predictive Models

Daniel Fink and Wesley Hochachka

**Abstract** The Hierarchical Predictive Model (HPM) is a semiparametric mixed model where the fixed effects are fit with a user-specified non-parametric component. This approach extends current spline-based semiparametric mixed model formulations, allowing for more flexible nonparametric estimation. Greater adaptability simplifies model specification making it easier to analyze data sets with large numbers of predictors. Greater automation also extends the scope of exploratory analyses that may be performed with mixed models. Using a HPM, the analyst may select the predictive model to best suit their needs, exploiting the strengths of currently available predictive methods. A simulation study is used to demonstrate the advantages of accounting for known hierarchical structure in predictive models and to illustrate the adaptability of current decision-tree based predictive models. A HPM of the relative abundance of the North American House Finch (*Carpodacus mexicanus*) is used to demonstrate exploratory analysis with a real data set.

## 1 Introduction

Hierarchical models have emerged as the preferred tool for analyzing large complicated data sets. Multifaceted processes can be factored into a series of simpler, conditionally independent sub-processes and a wide variety of parametric models can be incorporated. Bird monitoring data lend themselves to hierarchical treatment because data arise as the result of a stochastic observational process conditional on spatio-temporally varying biological processes. By separating these two processes, researchers have been able to address a number of important complications that arise in the analysis of ecological monitoring data. For example, parametric models have been developed to account for imperfect capture of species (e.g., Jolly 1965; Seber 1965; Amstrup et al. 2006), varying detection during gathering of observational data (MacKenzie et al. 2002; Gelfand et al. 2005), observer-specific effects such as mis-identifications or incorrect counts

---

D. Fink (✉)

Cornell Laboratory of Ornithology, 159 Sapsucker Woods Rd, Ithaca, NY 14850, USA  
e-mail: df36@cornell.edu

(e.g., Geissler and Sauer 1990; Thogmartin et al. 2004), and spatial correlation (Thogmartin et al. 2004; Wikle 2003; Wikle and Hooten 2006).

However, for many problems there is insufficient a priori knowledge to justifiably specify parametric models at all stages of the hierarchy. Often, it is not known which predictors should be included in a model. Even when important predictors have been identified, an appropriate functional form for their inclusion is unknown. In either case, the ability to specify a fully parametric model would still be desirable under many circumstances. Efficient exploratory tools are needed to discover patterns in data to account for and describe potentially complicated relationships between predictors and response. Exploratory analyses are an important means of hypothesis generation, and, ultimately, enable the specification of better parametric models.

Semiparametric models use as much parametric structure as is warranted by subject-area knowledge while relying on nonparametric techniques to automatically account for additional predictors and processes that are less well understood. This is a hybrid modeling strategy where the nonparametric components function as exploratory tools, automatically detecting and fitting patterns in the data while simultaneously taking into account parametric structure. Several successful semiparametric techniques have been built upon the Linear Mixed Model (LMM). These methods include the spline-based methods of Ruppert et al. (2003), Wood (2006) and Gu (2002). Each of these spline-based models can adapt over sums of smooth low-dimensional predictor functions while inheriting a well developed set of inferential tools from the LMM. Ideally, we would like to extend these methods to deal with large sets of predictors by utilizing nonparametric methods capable of automatically identifying important predictors and interactions, including high order interactions among predictors, and functional forms of relationships.

Within the last decade, data mining and machine learning techniques have emerged as some of the most successful tools for modeling complex, multi-dimensional data (Hand et al. 2001). These techniques are sophisticated nonparametric tools for data exploration with a focus on producing accurate predictions. Data mining methods include neural networks, decision trees, and support vector machines. Many of these methods have been gaining recognition within the ecological community (De'ath and Fabricius 2000; Elith et al. 2006; Hochachka et al. 2007). These methods are capable of sifting through large number of predictors to identify important ones, their interactions, and functional forms (Hastie et al. 2001). The weakness of these methods are their limited ability to incorporate prior information, especially patterns of correlation. Most current implementations of these tools assume independence among the data.

The purpose of this article is to develop a modeling framework that combines the complementary strengths of the LMM and modern nonparametric predictive models. We call this framework the Hierarchical Predictive Model (HPM). It is a semiparametric mixed model where the fixed effects are fit with a user-specified predictive model. We do this by fitting the HPM as a Bayesian model using a simple Gibbs sampler. The Gibbs sampler allows us to iteratively update fixed and random

effects from nonparametric and parametric models, respectively. Taking an Empirical Bayes approach, we estimate the conditional fixed effects with the nonparametric predictive model. Thus, a wide variety of predictive models may be used to explore the fixed effects. Using data mining and machine learning methods, the HPM extends current spline-based semiparametric formulations, allowing for more flexible nonparametric estimation. Greater adaptability simplifies model specification making it easier to analyze data sets with large numbers of predictors. This will be of increasing importance as larger data sets become available to ecologists.

As a hybrid methodology, the HPM draws upon several other modeling frameworks. In Section 2 we review these modeling frameworks. The HPM and its fitting algorithm are developed in Section 3. In Section 4 we present results from a simulation study to demonstrate the advantages of accounting for known hierarchical structure into predictive models and to illustrate the adaptability of current decision-tree based predictive models. In Section 5, a HPM of the relative abundance of the North American House Finch is used to demonstrate an exploratory analysis with a real data set. We conclude with a brief discussion.

## 2 The Models

In this section we review several modeling frameworks developed for correlated and uncorrelated observations. Emphasis is placed on brief descriptions of the model frameworks noting their scope of application, strengths, and limitations. We distinguish between parametric and nonparametric frameworks and note how their strengths make them well suited for confirmatory and exploratory analysis, respectively.

### 2.1 Predictive Models

We use the term “predictive model” to refer to any model that extracts information from a set of predictors and independent responses to make future predictions. Let  $y_i$ ,  $i = 1, \dots, N$  be the responses each associated with  $p$  predictors  $x_i = [x_{1,i}, \dots, x_{p,i}]$ . It is assumed that each observation,  $y_i$ , arises as an independent realization from some true but unknown function,  $F(x_i)$  that maps  $x_i$  to  $y_i$ . The goal of predictive modeling is to use the data to estimate  $F(x)$  while minimizing the expected value of some specified loss function. Predictive models have been developed in various disciplines with their own unique sets of terminology. In statistics, this predictive problem is known as regression. In the machine learning and data mining communities it is known as the supervised learning problem and the term “regression” refers more specifically to supervised learning problems with a continuous response.

In this paper we will restrict our attention to normally distributed observations and write  $y = F(X) + \varepsilon$  where  $y$  is the  $n \times 1$  vector of observations and  $X$  is the  $n$

$\times p$  design matrix of predictors. The  $n \times I$  error vector,  $\varepsilon$ , is assumed arise from an uncorrelated normal distribution with zero mean and variance  $\sigma^2$ .

### 2.1.1 Parametric Predictive Models

Parametric models have an explicit parametric form where model parameters describe a known or hypothesized process of interest. For example, consider the classic normal linear model in statistics,  $y = X\beta + \varepsilon$ , where  $\beta$  is a  $p \times I$  vector of parameters. The conditional mean of  $y$  is modeled parametrically as a linear combination of predictor effects. Parametric modeling requires enough knowledge about the process being investigated to specify the model. Constructing good parametric models can take considerable time and effort. The strengths of parametric models are the ease of interpretation and the availability of inferential tools. A well developed body of statistical methods, both Frequentist and Bayesian, can be used to make inferences about the parameters of interest and about predicted observations. For this reason, parametric models are most often used to make confirmatory inferences.

### 2.1.2 Nonparametric Predictive Models

Nonparametric models are often described as models without parameters or without parameters of direct inferential interest. Here, we use the term “nonparametric” to describe predictive models that automatically adapt to patterns in data – this being the essential distinguishing quality. Adaptive models are designed to automatically *discover* patterns. This makes them especially well suited for exploratory analysis. The more adaptive the methodology, the greater the scope of exploration.

Generalized Additive Models (GAMs) (Buja et al. 1989; Hastie and Tibshirani 1990; Wood 2006) are a popular class of nonparametric statistical models for representing a response as the sum of low-dimensional smooth functions of predictors. The simple GAM:

$$y = f_1(x_1) + \varepsilon,$$

can be used to detect and describe nonlinear functional effects of  $x_1$ . This GAM can be extended to simultaneously estimate smooth joint effects of  $x_2$  and  $x_3$  by adding an appropriate term, like a tensor plate spline, to yield,

$$y = f_1(x_1) + f_2(x_2, x_3) + \varepsilon.$$

This GAM makes 3 assumptions about the systematic effects of the predictors on the response; (1) the functional effects  $f_1$  and  $f_2$  vary smoothly with the predictor values, (2) predictor  $x_1$  does not interact with  $x_2$  or  $x_3$ , and (3) predictors  $x_2$  and  $x_3$  are allowed to interact. This GAM can be used to detect if there is a 2-way

interaction among the user specified pair of predictors. Conceptually, one can extend this idea using higher order terms to automatically adapt to more complex multivariate functional forms.

Decision Trees were designed to automatically fit high-dimension multivariate functional forms. Using a strategy of binary recursive partitioning, these models adapt over high dimensional tensor-product predictor spaces to fit models with possibly high-order interactions. Thus, a decision tree model of the form

$$y = f(x_1, \dots, x_p) + \varepsilon$$

can be used to investigate numerous functional relationships. Predictive experiments can be used to extract information for identifying important predictors, describing their effects, and identifying interactions within sets of predictors; see Sections 4 and 5.

Nonparametric models vary widely in the type of adaptation they do and the strategies used to achieve them. Many highly-adaptive nonparametric predictive methods have been developed within the data mining and machine learning communities where problems are characterized by very large data sets, both in terms of the number of responses ( $N$ ) and the number of predictors ( $p$ ). Consequently, these methods are designed to be very efficient, both in terms of analyzing large numbers of responses as well as extracting predictive information from large sets of predictors. These methods include decision trees (e.g., Breiman et al. 1984), neural nets (e.g., Mitchell 1997), Support Vector Machines (SVMs) (e.g., Cristianini and Shawe-Taylor 2000), and ensemble variants of tree-based methods (e.g., bagged and boosted decision trees, random forests; e.g., Breiman 1996; Breiman 2001). Recently, these methods have enjoyed increasing visibility and application within the ecological literature, see De'ath and Fabricius (2000), Elith et al. (2006), and Hochachka et al. (2007).

## 2.2 Hierarchical Models

With hierarchical models, one can factor complicated, multifaceted processes into a series of simpler conditionally independent parametric sub-processes. When data have obvious hierarchical structure, it is advantageous to model this structure parametrically. The hierarchical model is a formal mechanism for pooling information from correlated responses, potentially making substantial improvements in model efficiency. In disciplines where correlated data are frequently confronted, specialized statistical models have been developed to deal with these correlations. For example, Kriging was developed for geo-statistical analysis where spatial correlation play is very important. Longitudinal analyses explicitly take into account the correlation induced by making several observations on individual experimental units over the duration of an experiment, e.g., patients in clinical trial.

Hierarchical models have also been developed to model a wide variety of processes including spatial data with varying support (Wikle and Berliner 2005;

Banerjee et al. 2004), measurement error models (Berry et al. 2002), dispersion processes (Wikle 2003) and dynamic processes (West and Harrison 1997; Banerjee et al. 2004).

In this paper we will focus on the Linear Mixed Model (LMM), a two-level parametric hierarchical model. The strength of this model is its success as a powerful framework within which to model patterns of correlation. By connecting LMMs together one may assemble more complex hierarchical structures and patterns of correlation, e.g., multilevel models (Goldstein 1995).

### 2.3 Linear Mixed Models

The Linear Mixed Model extends the linear model by incorporating random effects, which can be regarded as additional error terms, to account for correlations among observations. The general form of the LMM is

$$y = X\beta + Zu + \varepsilon$$

where  $y$  is a vector of  $N$  observable random variables,  $\beta$  is a vector of  $p$  unknown parameters having fixed values (fixed effects),  $X$  is the  $n \times p$  fixed effect design matrix, and  $Z$  is the  $n \times q$  random effect design matrix. Both  $u$  and  $\varepsilon$  are unobservable random vectors (random effects) of length  $q$  and  $n$ , respectively. We will refer to  $u$  to as the “random effects” and  $\varepsilon$  as the “error” term to distinguish them. It is assumed that both the random effects and errors are normally distributed and uncorrelated with each other. Specifically,  $u \sim N(0, \Sigma(\varphi))$  where  $\Sigma(\varphi)$  is assumed to be a parametric covariance model with variance component(s)  $\varphi$  and  $\varepsilon \sim N(0, \sigma^2 I)$  where  $\sigma^2$  is a positive constant and  $I$  is the  $n$ -dimensional identity matrix.

The LMM is one of the most useful models in modern statistics, allowing many complications to be handled within the familiar linear model framework. This model has become a standard approach to model genetic effects, longitudinal data, blocked designs, crossed designs, nested designs, varying coefficient models, and numerous problems with temporal and spatial correlation (see Robinson 1991; McCulloch and Searle 2001; Zhao et al. 2006 for good reviews). One of the reasons for the success of the LMM is the ease and efficiency with which correlation structure can be incorporated into the model. Often, a basic understanding of the correlation structure is sufficient knowledge to specify useful covariance models for the random effects. The vast literature on LMM is a testament to this fact.

Like any parametric model, the LMM requires enough *a priori* information to specify the entire model. For each process that is included in the model, the analyst must decide which predictors to include, which predictors interact, and the functional form of all effects. When there is more predictor information than prior knowledge, it may difficult to specify a good fixed effect model, ultimately limiting the amount of covariate information that can be admitted into the model. This becomes a bigger problem as the number of predictors grows and *a priori* information does

not increase proportionally. In practice, this limits the amount of information that may be brought to bear on confirmatory analyses and it is the reason that the LMM is not often used for exploratory analyses.

## 2.4 *Semiparametric Mixed Models*

Semiparametric predictive models incorporate flexible nonparametric model components within a parametric framework. This gives the analyst the ability to include as much parametric structure as can be justified by subject-area knowledge while using adaptive nonparametric components to automatically search for additional signal in the data. The use of this hybrid modeling strategy can improve confirmatory analysis by automatically incorporating additional predictor information, with fewer unjustified assumptions, than possible in traditional parametric models. Semiparametric models may also be used to conduct focused exploratory analyses by adaptively searching for patterns after accounting for known parametric structure.

The SemiParametric Mixed Model (SPMM) includes nonparametric model components to automatically incorporate fixed effect predictor information within the LMM framework. Extending the LMM of the previous section, we write the general SPMM as

$$y = f(X) + Zu + \varepsilon$$

where  $f(X)$  represents a nonparametric predictive component for fixed predictor effects.

Some of the most effective semiparametric modeling strategies to take advantage of the mixed model framework have been based on penalized splines. These approaches use spline basis-expansions as flexible function effects and then control the complexity of the fit by means of penalization. The key to incorporating penalized splines within the mixed model framework is to recast the penalty as a random effect. Current implementations differ in the types of spline functions and fitting strategies used. Current examples include the penalized regression splines of Ruppert et al. (2003), the generalized additive models of Wood (2006) and the smoothing spline ANOVA models of Gu (2002), though the connection between penalized spline methods and the mixed model has a much longer history (see Wahba 1990).

There are two main advantages to bringing this nonparametric smoothing technique to the mixed model. First, it allows splines to be used with a wide variety of data types and diverse applications where mixed models are already used. Second, it give practitioners access to many of the inferential tools developed for the mixed model. A serious limitation of this strategy is computational. In order to adapt to functional forms in high-dimensional spaces, it is necessary to generate very large spline-basis expansions which in turn require the manipulation of equally large

matrices. This is why most current techniques limit the response to be the sum of several low-dimensional smooth functions of predictors.

### 3 Hierarchical Predictive Models

The HPM, is a SPMM

$$y = f(X) + Zu + \varepsilon,$$

where  $y$  is a vector of  $N$  observable random variables,  $X$  is the  $n \times p$  fixed effect design matrix,  $Z$  is the  $n \times q$  random effects design matrix, and  $f(X)$  is a vector of  $N$  predictions. For notational convenience, we will denote the vector of fixed effects as  $f$ , suppressing its dependence on the predictors in the fixed effect design matrix. The random effects are normally distributed,  $u \sim N(0, \Sigma(\varphi))$  where  $\Sigma(\varphi)$  is assumed to be a parametric covariance model with variance component(s)  $\varphi$ . The errors are independent and normally distributed  $\varepsilon \sim N(0, \sigma^2 I)$  where  $\sigma^2$  is a positive constant and  $I$  is the  $n$ -dimensional identity matrix. The errors and random effects are assumed to be independent of each other.

Although hierarchical models are not inherently Bayesian, complex hierarchical models are most easily fit within the Bayesian framework using simulation-based Markov Chain Monte Carlo (MCMC) techniques. Bayesian inferences are based on the posterior distribution of the unknown model parameters conditioned on all observed, known quantities. The posterior distribution for the HPM is  $[f, u, \varphi, \sigma^2 | y, X, Z]$ . We denote the distribution of a random vector  $x$  by  $[x]$  and the conditional distribution of  $y$  given  $x$  is by  $[y|x]$ . The conditional dependence of posterior distributions on  $X$  and  $Z$  will be omitted for notational convenience. The MCMC sampler used to fit the Bayesian HPM is described below.

The Gibbs sampler (Robert and Casella 2004) is used to simulate the posterior by breaking the vector of model parameters into convenient subsets and iteratively sampling from the resulting conditional distributions. The hierarchical structure of the mixed model naturally breaks down into conditional distributions for  $u$ ,  $f$ , and the variance components  $\varphi$  and  $\sigma^2$ ,

$$\begin{aligned} & [u | f, \varphi, \sigma^2, y] \\ & [f | u, \varphi, \sigma^2, y] \\ & [\varphi | f, u, \sigma^2, y] \\ & [\sigma^2 | f, u, \varphi, y] \end{aligned}$$

The Gibbs sampler generates samples from each posterior conditional distribution to sequentially update the parameters. Strategies for updating the parameters vary depending on the form of the conditional distribution.

The conditional distribution of  $u$  is proportional to the product of normal distributions,

$$\begin{aligned}
 [u|f, \varphi, \sigma^2, y] &\propto [u|\Sigma(\varphi)][y|f, u, \sigma^2] \\
 &= \exp\left[-\frac{1}{2}u^T \Sigma^{-1}(\varphi)u\right] \exp\left[-\frac{\sigma^2}{2}(f + Zu)^T (f + Zu)\right].
 \end{aligned}$$

This distribution is conditionally conjugate, meaning that it has an analytically tractable form. In this case, the conjugate posterior is also normal (Lindley and Smith 1972), making it is straightforward to simulate. Most non-normal random effects will give rise to non-standard, analytically intractable full conditionals which require MCMC techniques.

Instead of sampling directly from the conditional distribution of  $f$ , our strategy is to use a predictive model to *estimate* the expected conditional fixed effects,  $\hat{f} = E[f|u, \varphi, \sigma^2, y]$ . These estimates are plugged into the Gibbs sampler to update  $f$ . Conditioning on the random effects,  $u$ , we consider  $Zu$  as an initial estimate of the predicted observations. This estimate can be improved by taking into account the systematic effects of the predictors,  $X$ . This is where we use the predictive model to estimate the expected responses  $\hat{f}$  by regressing the residuals  $r = y - Zu$  on predictors  $X$ .

The best strategy for sampling from the full conditionals of the variance components depends on the specific covariance model  $\Sigma(\varphi)$  and the prior distributions specified for  $\varphi$  and  $\sigma^2$ . For example, when  $\Sigma(\varphi) = \varphi I$ , as in repeat measures designs or the error term, the inverse gamma distribution is conditionally conjugate. Other prior specifications will require MCMC methods, e.g., reference priors (Zhao and Wells 2005). When the form of  $\Sigma(\varphi)$  is more complex, e.g., autoregressive (AR) processes or Matern covariance models, a general purpose algorithm like Metropolis-Hastings can be used to generate samples from the conditionals.

To summarize, the Gibbs sampling algorithm is:

1. Initialize MCMC parameters:  $u^{(0)}, \varphi^{(0)}, \sigma^2(0)$
2. For  $m$  in  $1$  to  $M$  do:
3. Predict  $f^{(m)}$  from the residuals  $r = y - Zu^{(m-1)}$  and covariates  $X$ 
  4. Sample random effects  $u^{(m)} \sim [u|f^{(m)}, \varphi^{(m-1)}, \sigma^{2(m-1)}, y]$
  5. Sample variance component  $\varphi^{(m)} \sim [\varphi|f^{(m)}, u^{(m)}, \sigma^{2(m-1)}, y]$
  6. Sample variance component  $\sigma^{2(m)} \sim [\sigma^2|f^{(m)}, u^{(m)}, \varphi^{(m)}, y]$
7. end For
8. end Algorithm.

Because we estimate the fixed effects,  $f$ , this algorithm is not, strictly speaking, Bayesian. Methods that replace unknown quantities with data-based estimates and then perform Bayesian analysis are known as “empirical Bayes”. Empirical Bayesian methods are often used because they allow the analyst to take advantage of prior information in a simplified way without having to specify prior distributions. The resulting empirical Bayes estimators often have good frequentist properties, though theoretical results have been established only for certain estimators (Lehmann and Casella 1999). One disadvantage of estimating parameters with the empirical Bayes approach is that the method does not account for the variability in

the estimation step. For this reason, we suggest that all confidence regions based on the HPM posterior be considered only approximate, and most likely biased small. Discussions of the this underestimation in posterior variance, along with remedies, can be found in Carlin and Louis (2000).

The Gibbs sampler can be started with initial values  $u^{(0)}$ ,  $\varphi^{(0)}$ ,  $\sigma^{2(0)}$  set equal to estimates from a LMM with the same random effects design and some reasonably simplified fixed effects model. The number of iterations,  $M$ , required for convergence to the stationary distribution depends on the complexity of the random effect design and the degree of correlation between fixed and random effects. With relatively simple random effects and little correlation between fixed and random effects, we have found that chains of several thousand iterations are sufficient to achieve convergence.

## 4 Simulation Study

The following simulation study demonstrates the potential advantages of accounting for known hierarchical structure with predictive models. The predictive performances of several decision-tree based models are compared when used on their own and when embedded within a HPM with known hierarchical structure. With the HPMs predictive power improves and functional structure may be fit and discovered. The posterior distribution of fixed-effect predictions is explored to illustrate the adaptability of decision-tree based predictive models.

### 4.1 Performance Comparisons with Dependent Data

The data for this simulation were constructed to include several functional features commonly found in ecological data. Dependence among the observations arise from two separate processes; spatial correlation that describes the similarity of neighboring observations and observer effects that describe the similarity among observations made by the same observer. The parametric hierarchical model used to generate the observations is

$$y = f(X) + Z_s u_s + Z_o u_o + \varepsilon,$$

where  $y$  is a vector of  $N$  observations. Observation errors  $\varepsilon$  are normally distributed conditionally independent on the process with variance  $\sigma^2 = 4$ . The fixed effects model is

$$f(X) = -4.5 + 5I(x_1 > 0.5) - 6x_4 + 2 \sin(6\pi x_6) + \frac{\sin(6\pi r)}{r},$$

where  $I(x_1 > 0.5)$  is the indicator function that takes on the value of 1 when  $x_1 > 0.5$  and zero otherwise and  $r = \sqrt{(x_9 - 0.5)^2 + (x_{10} - 0.5)^2}$ . This model includes a

threshold effect ( $x_1$ ), linear effect ( $x_4$ ), oscillating effect ( $x_6$ ), and a complex 2-way interaction between  $x_9$  and  $x_{10}$ .

The spatial effects  $u_s$  are modeled as a zero-mean, isotropic Gaussian process. Let  $u_s \sim N(0, \Sigma(s))$  where the covariance matrix  $\Sigma(s)$  describes the covariance between locations,  $s$ . The covariance between locations  $s_i$  and  $s_j$  decays exponentially as a function of the distance between them  $\Sigma_s(s_i, s_j | \rho, \sigma_s^2) = \sigma_s^2 \exp(-\|s_i - s_j\|/\rho)$  with range parameter  $\rho = 0.05$  and scale parameter  $\sigma_s^2 = 16$ . For computational convenience we assume that the range parameter is known.

In order to control the size of the spatial effects, and the computations necessary to handle them, we model the spatial correlation  $u_s$  as a  $50 \times 1$  vector of spatial effects at 50 selected “reference locations”. The resulting spatial covariance  $\Sigma(s)$  is a  $50 \times 50$  reduced rank correlation matrix (Ruppert et al. 2003, Section 13.4). Reference locations were determined as the centroids of the neighborhoods generated from a k-nearest neighbor analysis of the observation locations, reflecting the spatial density of the observations. The spatial design matrix  $Z_s$  is the corresponding  $N \times 50$  exponential covariance matrix between the observed locations and the reference locations. The spatial correlation among the  $N$  observations is calculated as the product  $Z_s u_s$ , similar to the Kriging prediction equations.

It is assumed that each observation was made by one of ten individual observers selected at random with equal probability. We further assume that each observer is biased and that the population of these biases or “observer effects”,  $u_o$ , are independent and normally distributed with variance  $\sigma_o^2 = 16$ . The observer effect design matrix is an  $N \times 10$  indicator matrix with elements  $\{Z_o\}_{i,j}$  equal 1 if the  $i$ -th observation was made by the  $j$ -th observer and 0 otherwise. Thus, the factor  $Z_o u_o$  induces correlation among observations made by the same observer.

Each simulated data set consists of  $N = 2000$  observed responses from the model specified above. A total of ten fixed effect predictors were generated of which only the 5 indicated above influence the response. Each predictor  $x_{i,j}$ ,  $i = 1, \dots, N$ ,  $j = 1, \dots, 10$  was generated independently on a uniform random distribution between 0 and 1,  $U[0, 1]$ , and stored in the  $N \times 10$  fixed effect design matrix,  $X$ . Locations  $s_i$ ,  $i = 1, \dots, N$  were generated randomly as independent latitude-longitude pairs on  $U[0, 1]$ , denoted as predictors  $x_{11}$  and  $x_{12}$ , respectively. Predictor  $x_{13}$  is the  $N \times 1$  vector of randomly generated labels for the ten observers. The signal-to-noise ratio is

$$\frac{\text{var}(f(X) + Z_s u_s + Z_o u_o)}{\sigma_\epsilon^2 + \sigma_s^2 + \sigma_o^2} \approx 1.24.$$

We compare the performance of four decision-tree methods. Decision trees, as a general class of models, have several features that make them a good choice of predictive model: (1) they are relatively easy to implement and understand, (2) they automatically discover and fit interactions including high-order interactions and (3) most implementations automatically impute missing predictor values. The simplest decision tree approach used here is the “rpart” model (Therneau and

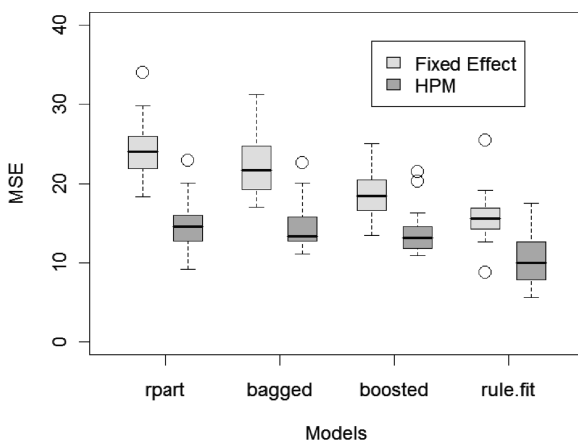
Atkinson 2007) which produces a single Classification and Regression Tree (CART) fit by cost complexity pruning (Breiman et al. 1984). In order to control the highly variable predictions of CART trees, Breiman (1996) suggested averaging predictions from a bootstrap sample of deliberately overfit CART trees. These “Bootstrap AGgregations” are known as “bagged” decision trees and usually outperform single trees. Boosting is another successful method used to average predictions across many simpler decision trees. It is equivalent to fitting an additive expansion in a set of basis functions (Hastie et al. 2001). We use the boosted decision trees implemented in the `gbm` library in R (Ridgeway 2006). RuleFit is another ensemble method that uses LASSO penalization (Tibshirani 1996) to combine predictions from individual trees (Friedman and Popescu 2005).

Each realization of the data was fit with all four decision-tree models and their corresponding HPMs. In order to make a fair comparison between the decision tree models and the HPMs we gave each of the decision-tree models access to the same predictor information utilized by the HPMs. Thus, each decision tree was fit using *all* 13 predictors including the latitude, longitude, and the vector of observer identifiers.

Model performance is measured as the Mean Squared Error (MSE) between the true and predicted responses. To guard against overfitting, the MSE is computed on an independent test set of data. All test predictions are made at new locations, for new observers so as to avoid any potential overfitting of the random effects, that is, overfitting location-specific or individual observer effects. The LMM BLUP estimator is  $\hat{y} = \hat{f}(X) + Z_P \hat{u}_s$ , where “hats” denote estimates and  $Z_P$  is the covariance between the new locations,  $s_i$   $i = 1, \dots, 1000$  and the reference locations  $s_j$   $j = 1, \dots, 50$ ,  $\{Z_P\}_{i,j} = \Sigma_s (s_i, s_j | \rho, \sigma_s^2)$ . The HPM predictions use the mean marginal posterior estimates for the fixed effects, variance components, and spatial effects. For the decision tree models, we “average out” estimated observer-specific biases by computing the mean predicted response where the mean is taken over the set of observers in the data set used for model training.

Half of each data realization was randomly assigned to training and testing sets. A single training-test set was used, instead of k-fold cross validation to expedite calculations. The simulation study was based on 100 trials. Diffuse Inverse Gamma (IG) priors were used for all the variance components,  $[\sigma^2] = [\sigma_s^2] = [\sigma_o^2] = \text{IG}(a = 0.1, b = 0.1)$ . MCMC chains were initiated with true values to reduce computations time. Each chain was run for 1000 iterations. All computations in this paper were performed with the R statistical computing language (R Development Core Team 2006).

Boxplots of the test set MSE are shown in Fig. 1. The variation in MSEs is due to the Monte Carlo error, estimate and model uncertainty, and variation from the test-train split. The mean square error is seen to vary among the decision tree models with the largest errors for `rpart` and smaller errors for each of the ensemble methods. The performance of all decision tree models improves when the methods are embedded in the hierarchical model. The HPM based on RuleFit was the best overall performer. These results suggest the kinds of performance gains possible when covariance patterns exist and are correctly modeled in the hierarchy rather than modeled nonparametrically as fixed effects.



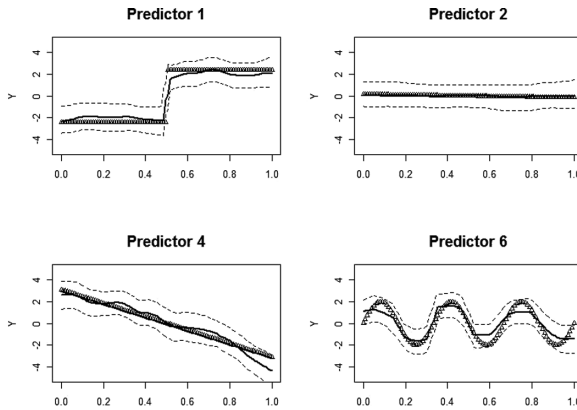
**Fig. 1** The test set Mean Squared Error (MSE) between the true responses and the predicted responses are shown for Decision Tree models (Fixed Effect) and HPMs, organized by the type decision tree model. The boxplots show the variation in MSEs due to Monte Carlo error, estimate and model uncertainty, and variation from the test-train split. The performance of all decision tree models improves when the methods are embedded in the hierarchical model

### 4.2 Partial Dependence Plots for Effect Exploration

Although nonparametric predictive models have good predictive performance, many are essentially “black box” methods, making them difficult to interpret. The same is true of the HPM where all fixed effect information is stored as a high-dimensional joint posterior distribution of predictions. Partial dependence functions (Friedman 2001; Hastie et al. 2001; Hooker 2007; Hochachka et al. 2007) are a simple general purpose tool for visualizing and exploring predictor effects. We use partial dependence plots to explore a fixed-effects posterior distribution and use these plots to illustrate the adaptability of RuleFit within the HPM.

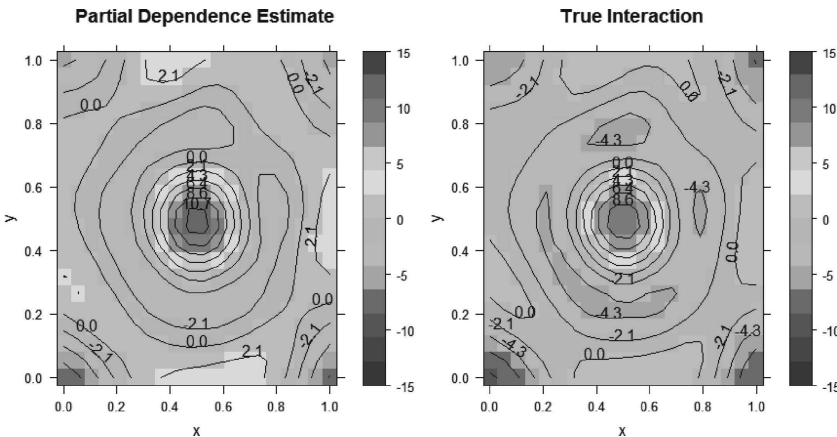
We begin by investigating the effects of each of the 10 individual fixed effect predictors from a single data realization. A natural approach to this investigation is to plot the predictions as a function of a single predictor. Unfortunately, the resulting trend may be simultaneously affected by any number of predictors that affect the response, making it difficult to isolate and describe the effects of any individual predictor. In order to better isolate the effect of each individual predictor we compute the effect of the predictor on the modeled response after accounting for the average effect of all other predictors. This is done by marginalizing over the joint distribution of all other predictors. These are one dimensional partial dependence plots. They best represent the effect of an individual predictor on the predicted response when the predictor’s effects are nearly additive. All partial dependence plots are centered at zero.

Univariate partial dependence plots of the posterior conditional means and approximate pointwise 90% Bayesian confidence regions for predictors  $x_1$ ,  $x_2$ ,  $x_4$ , and  $x_6$  are shown in Fig. 2. The partial dependencies for each mean effect were calculated at 100 equally spaced locations along the  $x$ -axis. Linear interpolations were plotted for the effects and confidence bounds. The approximate pointwise



**Fig. 2** Univariate partial dependence plots of the posterior conditional means and approximate pointwise 90% Bayesian confidence regions for the predictors  $x_1$ ,  $x_2$ ,  $x_4$ , and  $x_6$  (solid and dashed, respectively) and true effects (triangles) used in the simulation model described in Section 4.1

confidence regions were estimated with the 95th and 5th quantiles of the posterior partial effects. The RuleFit estimates of the posterior mean effects capture the main features of the true effects for all four predictors. Because of the discrete support of the DT basis, RF is also able estimate the sharp threshold in  $x_1$ . RuleFit’s penalization strategy may produce smooth effects like the oscillations in  $x_6$ . Predictor  $x_2$  is correctly identified as uninformative. The other four predictors were also identified as uninformative.



**Fig. 3** The two-dimensional partial dependence plots and true interaction surface for  $x_9 - x_{10}$  used in the simulation model described in Section 4.1. The interaction between these two predictors and their joint functional form were automatically detected and estimated by the predictive model, RuleFit

We calculated the two-dimensional partial dependence plots to investigate the RuleFit estimate of the  $x_9 - x_{10}$  interaction surface, Fig. 3 (left). Partial dependence estimates were made over a  $20 \times 20$  grid on the  $x_9 - x_{10}$  unit square and interpolated using a penalized spline. The true interaction surface, Fig. 3 (right), was evaluated and smoothed on the same grid. Contours and shading are the same for both panels to facilitate comparison. The strong similarity between these plots confirm RuleFit's ability to detect and fit complex interactions within HPM. RuleFit has automatically determined which predictors are additive and which interact. This makes RuleFit a good tool for automatically detecting interactions.

## 5 HPM Exploration of House Finch Abundance

The goal of this section is to demonstrate how the HPM can be used to explore patterns in real data. The discussion is presented at a conceptual level focusing more on analysis techniques and interpretation than the biological results. For this reason, we have deliberately chosen a model based on a simple hierarchical structure from a well understood species. We use HPM to model the relative abundance of North American House Finch at back yard feeders using data from the citizen-science winter monitoring program, Project FeederWatch (PFW, <http://www.birds.cornell.edu/pfw/>). The HPM has a random feeder-effect and a large set of previously unused predictors. The exploratory analysis is used to identify important new predictors and estimate some functional effects.

### 5.1 The Data

PFW is a winter-long "citizen science" monitoring project in which members of the general public throughout the United States and Canada record the maximum number of birds seen together, for each of the bird species that they see at their bird feeders. Observation periods occur over two consecutive days, at weekly or biweekly intervals. The program begins in mid-November and runs till the beginning of April. Participants record the location, date, bird numbers and effort expended during each observation period. They are also asked to provide data describing the weather and the environments around their feeder locations, such as presence or absence of coniferous and deciduous trees, water bodies, and the degree to which landscapes are altered by humans. Information is recorded about factors that may attract or deter nearby birds from being observed at a feeder such as the types of feed available, the number and configuration of the feeders, and the presence of pets and squirrels.

In addition to the information provided by PFW participants, we acquired several other descriptors of sites from the Avian Knowledge Network (see <http://www.avianknowledge.net/content/>) including descriptions of the general biogeographic region, local habitat, elevation, and human population density. These data were extracted based on the latitudes and longitudes of the PFW feeder sites. The complete data set included a total of 76 predictors, see Table 1.

**Table 1** Fixed effect predictors in HPM analysis used these 76 predictors. Seventy-two of the predictors were reported by PFW participants plus each site's Bird Conservation Region (BCR, see <http://www.nabci-us.org/map.html>), U.S. Census Bureau census block-level human population density estimate from 2000, elevation (2 from different digital elevation data sources and resolutions: USGS National Elevation Dataset, 10 m resolution data <http://www.mapmart.com/DEM/DEM.htm>; and GTOPO30, 30 arcsec resolution data <http://edc.usgs.gov/products/elevation/gtopo30/gtopo30.html>), and habitat type from the U.S. National Land Cover Database (NLCD) recorded as (one of 9 separate Anderson level 1 habitat classification categories within the grid block of the count site; U.S. National Landcover Data, 1992 version)

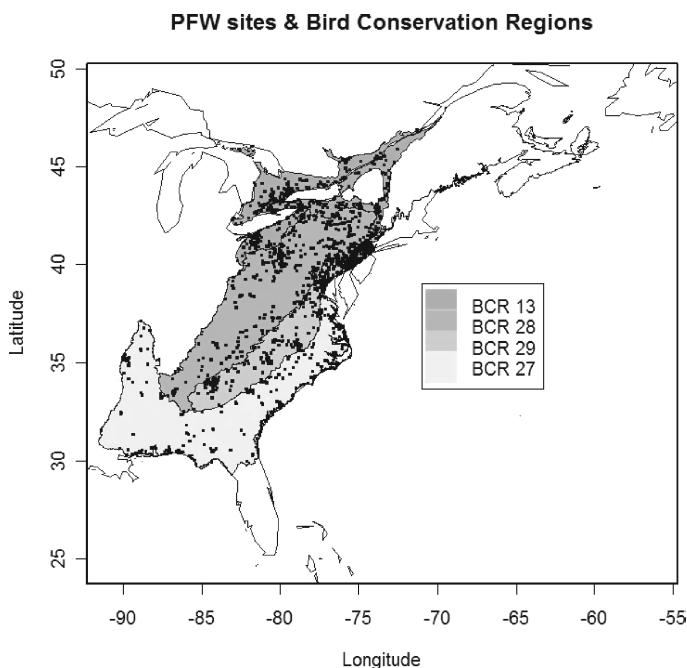
Temporal(2)	Attraction & deterrence to feeders (29)	Local habitat (31)
Season	count_area_size	nlcd
Date	fed_yr_round	yard_type_garden
	fed_in_jan	yard_type_landsca
	fed_in_feb	yard_type_woods
<b>Effort (2)</b>	fed_in_mar	yard_type_desert
	fed_in_apr	yard_type_pavement
	fed_in_may	hab_dcid_woods
	fed_in_jun	hab_evgr_woods
Effort 1	fed_in_jul	hab_mixed_woods
Effort 2	fed_in_aug	hab_orchard
	fed_in_sep	hab_park
	fed_in_oct	hab_water_fresh
	fed_in_nov	hab_water_salt
<b>Human Population Density (1)</b>	fed_in_dec	hab_residential
	numfeeders_suet	hab_industrial
Human.pop.density	numfeeders_ground	hab_agricultural
	numfeeders_hanging	hab_desert_scrub
	numfeeders_platfrm	hab_young_woods
	numfeeders_humming	hab_swamp
<b>Weather (6)</b>	numfeeders_water	hab_marsh
	numfeeders_thistle	hab_other
	numfeeders_fruit	evgr_trees_atleast
	bird_baths_atleast	evgr_shrbs_atleast
	high_feeders	dcid_trees_atleast
	nearby_feeders	dcid_shrbs_atleast
temp_lo	squirrels	fru_trees_atleast
temp_hi	cats	cacti_atleast
snow_coverage	dogs	brsh_piles_atleast
snow_depth	humans	water_srcs_atleast
snow_crusty		evgr_any_atleast
precipitation		dcid_any_atleast
<b>Physiographic (5)</b>		
latitude		
longitude		
elevation (categorical)		
elevation_1		
elevation_2		

Although the PFW data set contains a large number of potentially informative predictors, most of them have never been used to model the distribution or relative abundance of backyard species (Lepage and Francis 2002; Wells et al. 1998; Hochachka and Dhondt 2000; Hochachka and Dhondt 2006). Currently, there is

insufficient landscape-level understanding of the necessary ecological processes to specify a parametric model for all available predictors, or even a large subset of them. Such a task would require substantial exploration to determine (1) which predictors to include in the model, as well as (2) the functional form of the predictors, and (3) their interactions.

Missing data are another serious data complication that often affects how data are modeled. For example, of the 11,066 observations for NABCI Bird Conservation Region 13 (Lower Great Lakes/Saint Lawrence Plain; <http://www.nabci-us.org/map.html>), 9850 were missing at least one of the 72 PFW predictors — 89% of the records were incomplete. The expedient solution to the missing data problem is to throw out responses and/or predictors with missing data, but this reduces the information available for analyses and may introduce bias and increase the variance of results. More rigorous imputation options are often far more difficult to implement and require additional assumptions. For these reasons, many analyses are based on only a subset of the available data.

We analyzed the magnitude of positive group sizes from the 1993–1994 to the 2003–2004 seasons within the eastern North American range of the House Finch (Fig. 4). This species is well understood and has been independently analyzed in



**Fig. 4** This map shows the Project Feeder Watch (PFW) feeder sites as black squares within each of the four BCRs analyzed. BCR 13 is the Lower Great Lakes/St. Lawrence Plain region, BCR 28 is the Appalachian Mountains region, BCR 29 is the Piedmont region, and BCR 27 is the Southeast Coastal Plain region

the literature using PFW data (Hochachka and Dhondt 2006) providing a basis for validation. The spatial domain consists of four distinct Bird Conservation Regions. In order to simplify the comparison of regional variation in species' winter distributions, analyses were conducted separately for each of four different BCRs. For each BCR, a random sample of up to 400 "frequently participating" feeder sites were selected for analysis. "Frequently participating" feeders were defined to be sites that contributed at least 15 reports over the 11 season study period. The top 1% of maximum group sizes were trimmed to focus inference on smaller abundances by limiting the influence of the largest observations. Sample sizes were 11066, 5321, 12064, 11701 for BCRs 13, 27, 28, and 29 respectively. Four hundred locations were used for each BCR except BCR 27 which had 232.

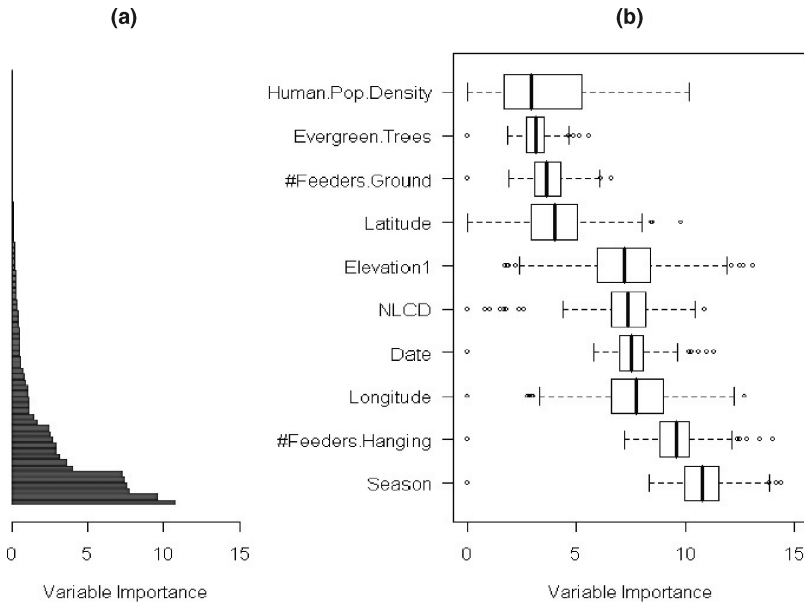
## 5.2 The Model

We use a mixed model with random feeder effects to account for correlation among observations from the same feeder. Let  $y_i$ ,  $i = 1, \dots, N$  be the natural log of the observed maximum count. We transform to the log scale to model errors as conditionally independent, additive normal noise  $y = f(X) + Zu + \varepsilon$ , as in Section 3. Random effects  $u$  estimate systematic differences among the  $q$  feeder locations. They are normally distributed  $[u|\sigma_f^2] = \text{Normal}(0, \sigma_f^2 I)$  where  $\sigma_f^2$  describes the amount of variation among the sites and  $I$  is the  $q$  dimensional identity matrix. The feeder effect design matrix  $Z$  is a  $N \times q$  indicator matrix with elements  $\{Z\}_{i,j}$  equal 1 if the  $i$ -th observation was made at the  $j$ -th feeder and 0 otherwise. We use RuleFit (Section 4.1) to estimate the HPM fixed effect,  $f(X)$ . The fixed effect design  $X$  is the  $N \times 76$  matrix of the predictors.

The data from each BCR were fit separately using diffuse inverse gamma (IG) priors for the variance components,  $\sigma^2 \sim \text{IG}(a = 0.1, b = 0.1)$  and  $\sigma_f^2 \sim \text{IG}(a = 0.1, b = 0.1)$ . Initial values for fixed effects were estimated using RuleFit and the residuals from this fit were used to initialize  $u$ . Realizations of  $(\sigma^2, \sigma_f^2, u, f)$  along with variable importances and partial dependences were collected on each iteration of the Gibbs sampler. Due to the simple structure of the hierarchy, each Gibbs sampler was expected to reach convergence quickly. The Raftery and Lewis (1992) diagnostic estimated that convergence sufficient to estimate the 5th percentile to within 1% accuracy with probability of 0.95 would be achieved with 1825 iterations. We computed 2600 iterations and discarded a burn-in of 100.

## 5.3 Exploratory Results

A first step towards uncovering the signal detected by the HPM is to rank the relative importance of its predictors. RuleFit computes a measure of relative variable importance designed to identify those variables that are used in its most influential predictive rules (see Section 7, Friedman and Popescu 2005). Relative importances



**Fig. 5** These plots show the relative variable importances for fixed effect predictors. The barchart in (a) shows the ordered marginal posterior median relative variable importance score for all 76 predictors. The boxplots (b) of the marginal posterior distributions for the relative variable importances of the 10 most important predictors, ordered according to their medians

are scaled to sum to 100 with larger values representing more important predictors. We collected the vector of relative variable importance at each Gibbs iteration. Figure 5a shows the barchart of the ordered marginal posterior medians for all 76 predictors. The exponential decay in importance is common among data sets with large numbers of predictors. Most information is concentrated among a small set of predictors.

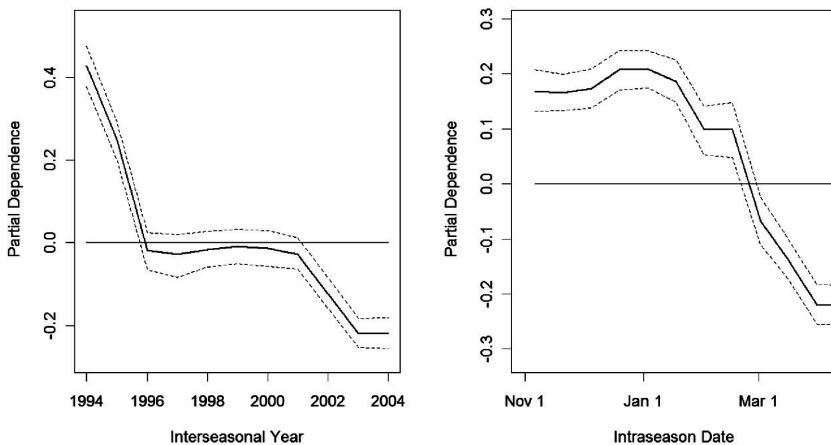
Figure 5b shows Boxplots of the marginal posterior distributions for the 10 most important predictors ordered according to their medians. This group of predictors captures many sources of variation known to be important for PFW data and for this House Finch population. The *Latitude–Longitude* pair describe the location of each feeder. *Season* is a 11-level ordinal predictor denoting the 11 different winter seasons. This is consistent with known changes in population trend over this time period due to the emergence of a novel bacterial pathogen, *Mycoplasma gallisepticum* (e.g., Dhondt et al. 2005). *Date* is a continuous Julian date starting at 1 on November 1st running to 150 on April 1st. This predictor could account for the known partial winter migrations of this House Finch population, as well as seasonal variation in propensity of the birds to visit feeders. RuleFit also identified two factors that attract birds to feeders as important. The *Number of Hanging Feeders* and the *Number of Ground Feeders* were ranked 2nd and 7th respectively. Three local habitat variables were ranked among the top 10 predictors. Landcover classification

(*NLCD*), elevation, and the presence or absence of evergreen trees, were ranked 5th, 6th, and 9th, respectively. *Human.Pop.Density* is human population density at the feeder location. This predictor may describe the association of House Finches with humans in suburban environments. Two observer effort variables describing the total hours of observer effort over each consecutive two day observation period and a 4-level ordinal variable that measures how many “halfday” periods were used for observation were also ranked highly, 13th and 17th, respectively.

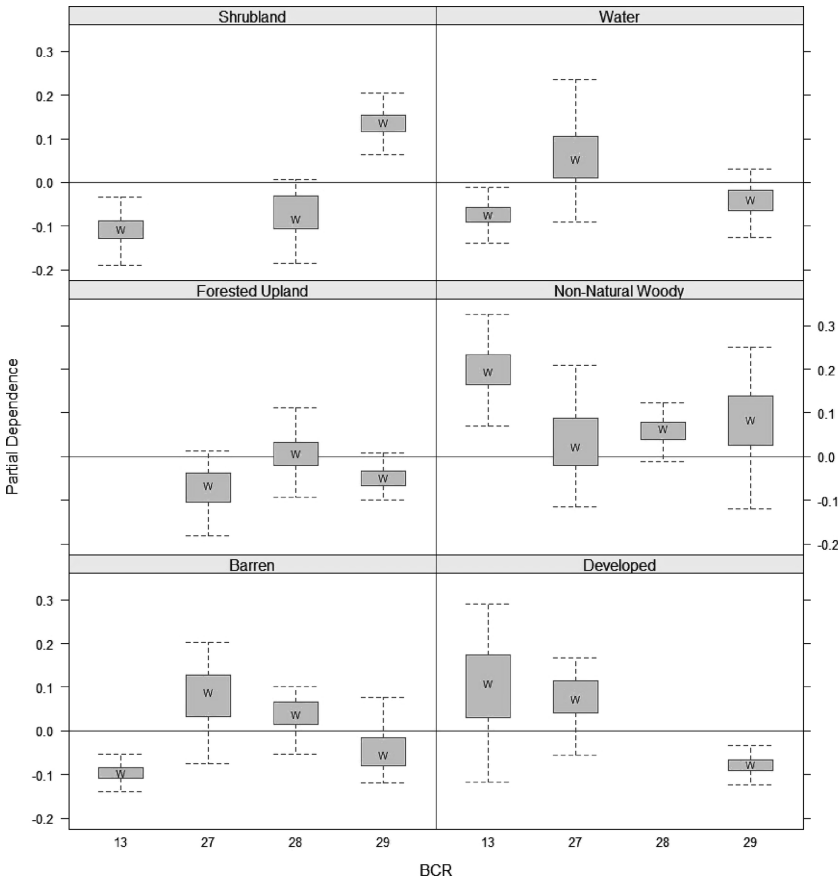
A benefit of a Bayesian analysis is the ability to produce estimates of uncertainty. These estimates incorporate the uncertainty from estimation and model uncertainty at the fixed effect level. The model uncertainty comes from the predictive model adaptation over complex model spaces. The boxplots of the marginal posterior variable importances show considerable uncertainty. Note that some of this uncertainty also arises because of predictor multicollinearity.

Partial dependence plots are used to visualize particular predictors effects. Figure 6 shows the partial dependence plots of the posterior conditional means and approximate 90% confidence regions for intra- and inter-seasonal trends in BCR 13. The inter-seasonal trend agrees with our expectation of a decline (Dhondt et al. 2005).

We also used partial dependence plots to explore how *NLCD* landcover effects vary by region, or BCR. For each BCR, we computed the partial dependence among the 9 *NLCD* Anderson level-1 land cover classes. Boxplots of the posterior partial dependence effect estimates were plotted for the six landcovers which were represented by at least three of the four BCR's, Fig. 7. The partial dependence among



**Fig. 6** Partial dependence plots for Intra- and Inter-seasonal trends in BCR 13 with pointwise 90% HPD confidence regions. The automatically detected intra-seasonal trends shown here agree with known changes in population trend over this time period due to the emergence of a novel bacterial pathogen, *Mycoplasma gallisepticum* (e.g., Dhondt et al. 2005). The inter-seasonal trend could account for the known partial winter migrations of this House Finch population, as well as seasonal variation in propensity of the birds to visit feeders.



**Fig. 7** The partial dependence among the NLCD classes are shown here with boxplots of posterior partial dependence among BCRs plotted by NLCD Anderson level-1 landcover classes. Boxes are centered *within* each BCR, facilitating the comparison of landcover effects within that BCR. However, to compare the *relative* effects of each landcover class across BCRs we grouped these relative effects by landcover. Substantial differences in the relative effects of several landcover classes across different BCR's suggest that local effects of land cover on group size vary regionally. For example, the observed group size of House Finches was smaller on shrublands, compared to other landcover classes, in BCR 13 and 28 while observed group sizes were larger on shrublands in BCR 29. The center "w" represents the median posterior partial dependence. Only landcover classes which were represented by at least three of our four BCR's have been plotted

the NLCD classes are centered *within* each BCR, facilitating the comparison of landcover effects within that BCR. However, to compare the *relative* effects of each landcover class across BCRs we grouped these relative effects by landcover. Substantial differences in the relative effects of several landcover classes across different BCR's suggest that local effects of land cover on group size vary among regions. For example, the observed group size of House Finches was smaller on shrublands, compared to other landcover classes, in BCRs 13 and 28 while observed group sizes were larger on shrublands in BCR 29.

This exploration suggests that habitat effects may operate over two distinct scales – the regional scale represented by BCR and a local scale of the  $30 \times 30$  m resolution NLCD layers. However, using these data it is not possible to tell if the habitat affects the probability to detection or the ecological process that governs abundance, or both. Interpretations of these effects are also confounded by multicollinearity, especially with other spatial predictors. The use of observational data may introduce sampling bias. For example, a known limitation of PFW data is its spatial footprint, which is concentrated towards anthropogenic habitats. Additional, carefully collected data would be needed to untangle and confirm the causes of the patterns that emerged from our analysis.

## 6 Discussion

In this article we have developed a highly adaptive semiparametric model by harnessing the complementary strengths of hierarchical and predictive models. From a mixed-model perspective, the strengths of the HPM are its great automation and adaptability. Large amounts of predictor information can be conveniently and quickly included in an analysis and explored.

While HPMs inherit many strengths from their parent models, they also inherit weaknesses. The parametric hierarchy itself must be specified by the analyst and this means that there will be a risk of misspecification. For many important problems patterns of correlation can be specified *a priori* with confidence, e.g., spatial analyses or repeated measures. In many hierarchies, the risk of misspecification may be mitigated by the adaptability of the hierarchical structure itself. For example, in several common LMM models, when there is insufficient evidence of variation among the random effects, the estimated variance components will take on limiting values that tend to “flatten” the hierarchy, effectively limiting the effect of misspecification.

Like any other model, multicollinearity among predictors makes it difficult to separately identify relationships between the correlated predictors and the response. This is an especially important challenge when exploring large environmental data sets where it is not uncommon to find large sets of multicollinear predictors. Finally, because HPMs require the computation of both the LMM and a predictive model, they can be computationally intensive.

It is important to remember that the data-mining component of an HPM does not carry the negative inferential properties of “data dredging”, a term that unfortunately is often viewed as synonymous with “data mining”. Chatfield (1995) defined the practice of data dredging as when an “analyst goes to great lengths to obtain a good fit. When a model is formed, fitted, and checked with the same data set in an iterative, interactive way”. Within the machine learning community there is a strong insistence on the use of independent data for testing and validation to guard against overfitting and dredging. Indeed, by performing exploration and regression in a single procedure, HPM actually avoids many common “dredging” problems that arise in more traditional multi-step approaches to regression model development and testing.

Because of their relative strengths and weaknesses, we view HPMs as serving three distinct roles in the analysis of observational data:

1. Where accurate predictions are the desired product of an analysis, HPMs are a logical class of models to use, because of their ability to make use of available information from the predictor variables, both when the forms of structural relationships are known and also where these things are unknown.
2. While many “products” from analysis of ecological data can be viewed as hypothesis validations, and there is good reason to conduct hypothesis-driven analysis of data, the specification of realistic and useful hypotheses requires prior knowledge. Where such knowledge does not exist, more exploratory analyses are appropriate, and efficient exploratory analyses will lead to creation of appropriate hypotheses more quickly. HPMs are particularly suited for such exploratory (hypothesis-generating) analyses when there is some amount of prior information available, as this prior knowledge can be incorporated into the exploratory model-building work.
3. Even when an analyst believes that (s)he has sufficient prior knowledge to specify an accurate parametric model, this is still merely a faith-based assertion unless there is some objective way of validating the appropriateness of a parametric model. HPMs can be used to assess the validity of fixed effect components within a hierarchical model by replacing them with more flexible nonparametric components and then comparing the overall predictive performances. Additional information may be gleaned from such a comparison by using partial dependence functions to compare the functional form and interactions of specific predictor effects estimated under both models.

**Acknowledgments** We thank the Avian Knowledge Network (AKN) team, Marty Wells, Charles Francis, and two anonymous reviewers for many helpful discussions and insightful comments. We would also like to thank the participants of Project FeederWatch and the staff in the Information Sciences unit at the Cornell Laboratory of Ornithology for their work in gathering and maintaining the data used in Section 5. Work on this paper was funded under the NSF Information Technology Research (ITR) for National Priorities program (award #EF-0427914).

## References

- Amstrup S, MacDonald L, Manly B (2006) *Handbook of Rapture-Recapture Analysis*. Princeton University Press, Englewood Cliffs, NJ 296 pp.
- Banerjee S, Carlin BP, Gelfand AE (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, London BocaRadon, FL 472 pp.
- Berry S, Carroll R, Ruppert D (2002) Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association* 97:160–169.
- Breiman L (1996) Bagging predictors. *Machine Learning* 24:123–140.
- Breiman L (2001) Random forests. *Machine Learning* 45:5–32.
- Breiman L, Friedman JH, Olshen RA, Stone JC (1984) *Classification and Regression Trees*. Chapman & Hall, New York.
- Buja A, Hastie T, Tibshirani R (1989) Linear smoothers and the additive model. *The Annals of Statistics* 17:453–555.

- Carlin BP, Louis TA (2000) Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall/CRC, Boca Raton, FL.
- Chatfield C (1995) Model uncertainty, data mining, and statistical inference. *Journal of the Royal Statistical Society, Series A*, 158:419–466.
- Cristianini N, Shawe-Taylor J (2000) An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge, UK.
- De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81:3178–3192.
- Dhondt AA, Altizer S, Cooch EG, Davis AK, Dobson A, Driscoll MJL, Hartup BK, Hawley DM, Hochachka WM, Hosseini PR, Jennelle CS, Kollias GV, Ley DH, Swarthout ECH, Sydenstricker KV (2005) Dynamics of a novel pathogen in an avian host: mycoplasmal conjunctivitis in House Finches. *Acta Tropica* 94(1):77–93.
- Elith J, Graham CH, Anderson RP, Dudik M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz G, Nakamura M, Nakazawa Y, Overton JMcC, Peterson AT, Phillips SJ, Richardson K, Scachetti-Pereira R, Schapire RE, Soberón J, Williams S, Wisz MS, Zimmermann NE (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29:129–151.
- Friedman JH, Popescu BE (2005) Predictive learning via rule ensembles. Technical Report, Stanford University.
- Geissler PH, Sauer JR (1990) Topics in route-regression analysis. In: Sauer JR, Droege S (eds) Survey Designs and Statistical Methods for the Estimation of Avian Population Trends. U.S. Fish and Wildlife Service, Biological Report 90(1):54–57.
- Gelfand A, Schmidt AM, Wu S, Silander JA, Latimer A, Rebelo AG (2005) Modelling species diversity through species level hierarchical modeling. *Applied Statistics*, 54:1–20.
- Goldstein H (1995) Multilevel Statistical Models. Halstead Press, New York..
- Gu C (2002) Smoothing Spline ANOVA Models. Springer, New York..
- Hand DJ, Mannila H, Smyth P (2001) Principles of Data Mining. MIT Press, Cambridge.
- Hastie T, Tibshirani R (1990) Generalized Additive Models. Chapman and Hall, London.
- Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Verlag, New York, 552 pp.
- Hochachka WM, Dhondt AA (2000) Density-dependent decline of host abundance resulting from a new infectious disease. *Proceedings of the National Academy of Sciences USA* 97:5303–5306.
- Hochachka WM, Dhondt AA (2006) House finch (*Carpodacus mexicanus*) population- and group-level responses to a bacterial disease. *Ornithological Monographs* 60:30–43.
- Hochachka WM, Caruana R, Fink D, Kelling S, Munson A, Riedewald M, Sorokina D (2007) Data mining for discovery of pattern and process in ecological systems. *Journal of Wildlife Management* 71(7):2427–2437.
- Hooker G (2007) Generalized functional ANOVA diagnostics for high dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics* 16(3).
- Jolly GM (1965) Explicit estimates from capture-recapture data with both death and immigration-stochastic models. *Biometrika* 52:225–247.
- Lepage D, Francis CM (2002) Do feeder counts reliably indicate bird population changes? 21 years of winter bird counts in Ontario, Canada. *Condor* 104:255–270.
- Lehmann E, Casella G (1999) Theory of Point Estimation, 2nd Edition. Springer-Verlag, New York.
- Lindley DV, Smith AFM (1972) Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 1–41.
- Mackenzie DI, Nichols JD, Lachman GB, Droege S, Royle JA, Langtimm CA (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology* 83(8):2248–2255.

- McCulloch CE, Searle SR (2001) *Generalized, Linear, and Mixed Models*. John Wiley and Sons, New York.
- Mitchell T (1997) *Machine Learning*. McGraw-Hill, New York.
- Raftery AE, Lewis SM (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science* 7:493–497.
- R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Ridgeway G (2006). gbm: Generalized Boosted Regression Models. R package version 1.5-7. <http://www.i-pensieri.com/gregr/gbm.shtml>.
- Robert CP, Casella G (2004) *Monte Carlo Statistical Methods*, 2nd Edition. Springer, New York.
- Robinson GK (1991) That BLUP is a good thing: the estimation of random effects. *Statistical Science* 8(1):15–51.
- Ruppert D, Wand MP, Carroll RJ (2003) *Semiparametric Regression*. Cambridge University Press, Cambridge, 402 pp.
- Seber GAF (1965) A note on the multiple-recapture census. *Biometrika* 52:249–259
- Therneau TM, Atkinson B, R port by Ripley B (2007) rpart: Recursive Partitioning. R package version 3.1-35. <http://mayoresearch.mayo.edu/mayo/research/biostat/splufunctions.cfm>
- Thogmartin WE, Sauer JR, Knutson MG (2004) A hierarchical spatial model of avian abundance with application to Cerulean Warblers. *Ecological Applications* 14(6):1766–1779.
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B* 58(1):267–288.
- Wahba G (1990) *Spline models for observational data* SIAM [Society for Industrial and Applied Mathematics] (Philadelphia).
- Wells JV, Rosenberg KV, Dunn EH, Tessaglia-Hymes DL, Dhondt AA (1998) Feeder counts as indicators of spatial and temporal variation in winter abundance of resident birds. *Journal of Field Ornithology* 69:577–586.
- West M, Harrison J (1997) *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York.
- Wikle CK (2003) Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology* 84:1382–1394.
- Wikle CK, Berliner ML (2005) Combining information across spatial scales. *Technometrics* 47:80–91.
- Wikle CK, Hooten MB (2006) Hierarchical bayesian spatio-temporal models for population spread. In: Clark JS and Gelfand A (eds) *Applications of Computational Statistics in the Environmental Sciences: Hierarchical Bayes and MCMC Methods*. Oxford University Press, Oxford.
- Wood SN (2006) *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, London/Boca Rarm, FL, 416 pp.
- Zhao X, Wells MT (2005) Reference priors for linear models with general covariance structures, Cornell Department of Statistical Sciences Technical Report.
- Zhao Y, Staudenmayer J, Coull BA, Wand MP (2006) General design Bayesian generalized linear mixed models. *Statistical Science* 21:35–51.