geographical analysis

*Geographical Analysis* (2016) **48,** 82–111

# Gaussian Process Regression and Bayesian Model Averaging: An Alternative Approach to Modeling Spatial Phenomena

Jacob Dearmon[1], Tony E. Smith[2]

[1]Department of Economics, Oklahoma City University, Oklahoma City, OK, USA, [2]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA

*Gaussian Process Regression (GPR) is a nonparametric technique that is capable of yielding reliable out-of-sample predictions in the presence of highly nonlinear unknown relationships between dependent and explanatory variables. But in terms of identifying relevant explanatory variables, this method is far less explicit about questions of statistical significance. In contrast, more traditional spatial econometric models, such as spatial autoregressive models or spatial error models, place rather strong prior restrictions on the functional form of relationships, but allow direct inference with respect to explanatory variables. In this article, we attempt to combine the best of both techniques by augmenting GPR with a Bayesian Model Averaging (BMA) component that allows for the identification of statistically relevant explanatory variables while retaining the predictive performance of GPR. In particular, GPR-BMA yields a posterior probability interpretation of model-inclusion frequencies that provides a natural measure of the statistical relevance of each variable. Moreover, while such frequencies offer no direct information about the signs of local marginal effects, it is shown that partial derivatives based on the mean GPR predictions do provide such information. We illustrate the additional insights made possible by this approach by applying GPR-BMA to a benchmark BMA data set involving potential determinants of cross-country economic growth. It is shown that localized marginal effects based on partial derivatives of mean GPR predictions yield additional insights into comparative growth effects across countries.*

## Introduction

Two of the most basic tasks of spatial statistical modeling are the *explanation* and *prediction* of spatial phenomena. As with all statistical modeling, the methods for achieving these goals differ to a certain degree. In spatial analyses, the task of explanation has focused mainly on parametric statistical models, typically some form of spatial regression, where identification of key variables can be accomplished by standard tests of hypotheses. But the need to specify prior

Correspondence: Jacob Dearmon, Meinders School of Business, Oklahoma City University e-mail: jdearmon@okcu.edu

Submitted: August 26, 2014. Revised version accepted: May 19, 2015.



doi: 10.1111/gean.12083

functional forms in these models tends to diminish their value for out-of-sample predictions. So the task of spatial prediction has focused on more flexible nonparametric approaches, typically local regression or stochastic interpolation methods.[1] But the very flexibility of these methods tends to impede the formal statistical identification of explanatory variables. Hence the objective of this article is to propose one method for unifying these two tasks. In particular, we combine a general form of stochastic interpolation known as *Gaussian Process Regression* (GPR) together with *Bayesian Model Averaging* (BMA).

Before doing so, it must be stressed that there have been other attempts to achieve such a unification. In the spatial literature, the work most closely related to our present GPR-BMA approach has been the efforts of LeSage and Parent (2007) and LeSage and Fisher (2008) to achieve more robust versions of the spatial errors model (SEM) and spatial autoregressive model (SAR) by combining them with BMA within a Markov Chain Monte Carlo (MCMC) framework.[2] More recently, Bivand, Gómez-Rubio, and Rue (2014) have discussed the use of the Integrated Laplace Approximation as a faster method to achieve marginal inference in spatial BMA routines. These spatial model averaging extensions, SEM-BMA and SAR-BMA, can in principle strengthen both the prediction and variable identification capabilities of spatial regression, and thus provide the natural benchmark for evaluating our present approach.[3]

For variable identification in GPR models, it should be stressed that this task can be accomplished in different ways. Perhaps the most widely known method is Automatic Relevance Determination (ARD), first introduced by Neal (1996) and MacKay (1998). While this method has great practical appeal, it offers little in the way of statistical identification of explanatory variables. Hence our present approach draws most heavily on the work of Chen and Wang (2010), who first employed BMA for both prediction and variable identification in GPR models (applied to the nonspatial problem of spectrometer calibration). The key feature of this approach is to allow uncertainties with respect to both relevant explanatory variables and predictions to be treated explicitly. In particular, GPR-BMA yields a posterior probability interpretation of simulated model-inclusion frequencies that provides a natural measure of the statistical relevance of each variable.

Within this framework, the main contributions of the present article are to develop this GPR-BMA model for spatial applications,[4] and in particular to show how it can be extended to analyze the localized marginal effects of spatial variables. The advantages of GPR-BMA for spatial analysis are then demonstrated both in terms simulated and empirical data sets. Using simulations, it will be shown that with essentially no prior knowledge of either functional forms or the nature of possible unobserved spatial autocorrelation, GPR-BMA is able to produce both reliable predictions and accurate identifications of relevant spatial variables. In contrast, both SEM-BMA and SAR-BMA are shown to be particularly sensitive to specification errors, even when spatial autocorrelation is captured exactly.

On the empirical side, we apply GPR-BMA to a data set that focuses on comparative economic growth between countries (Sala-i-Martin 1997; Fernandez, Ley, and Steel 2001a), and which has served as one of the standard benchmark data sets for BMA extensions of regression. To capture possible spatial effects, we include the spatial information for each country. In this context, it is shown that partial derivatives based on mean posterior GPR-BMA predictions allow the marginal impacts of relevant variables on growth rates to be localized by country, and used to draw comparative spatial inferences that are not available by more standard regression methods.

To develop these results, we begin with a detailed development of the GPR-BMA model. This is followed by selected simulation comparisons between GPR-BMA and the alternative BMA extensions of spatial regression. Finally, we develop our empirical application of GPR-BMA.

## GPR with BMA

In this section, we develop our proposed methodological procedure for spatial data analysis. This begins with a general development of Gaussian processes in a Bayesian setting that focuses on GPR—which amounts to posterior prediction within this framework. This is followed by a development of our BMA approach to GPR.

### Gaussian process regression

To set the stage for our present analysis, we start with some *random* (*response*) *variable*, $y$, which may depend on one or more components of a given vector, $x = (x_1, .., x_k)$, of *explanatory variables*, written as $y = y(x)$. If these explanatory variables are assumed to range over the measurable subset, $\mathcal{X} \subseteq \mathbb{R}^k$, then this relationship can be formalized as a stochastic process, $\{y(x) : x \in \mathcal{X}\}$, on $\mathcal{X}$. To study such relationships, the Bayesian strategy is to postulate a prior distribution for this process with as little structure as possible, and then to focus on posterior distributions of unobserved $y$-values derived from data observations. The most common approach to constructing prior distributions for stochastic processes, $\{y(x) : x \in \mathcal{X}\}$, is to adopt a *Gaussian Process* (*GP*) prior in which each finite subset of random variables, $\{y(x_1), .., y(x_N)\}$, is postulated to be multinormally distributed. In this way, the entire process can be specified in terms of a *mean function*, $\mu(x)$, and *covariance function*, $\text{cov}(x, x')$, $x, x' \in \mathcal{X}$, usually written more compactly as

$$y(x) \sim GP[\mu(x), \text{cov}(x, x')] \tag{1}$$

The simplest of these models assumes that the mean function is constant, and focuses primarily on relationships between variables in terms of their covariances. In particular, it is most commonly assumed that the mean function is zero, $\mu(x) = 0 \ , \quad x \in \mathcal{X}$, and that the covariance function has some specific parametric form, $\text{cov}(x, x') = c_\omega(x, x')$, designated as the *kernel function* for the process with (hyper)parameter vector, $\omega$. While there are many choices for kernels, one of the simplest and most popular is the *squared exponential kernel*,

$$c_\omega(x, x') = v \ \exp\left(-\frac{1}{2\tau^2} \ ||x - x'||^2\right) = v \ \exp\left[-\frac{1}{2\tau^2} \sum_{j=1}^{k} (x_j - x'_j)^2\right] \tag{2}$$

which involves two (positive) parameters, $\omega = (v, \tau)$. Hence all covariances are assumed to be positive, and to diminish as the (Euclidean) distance between explanatory vectors, $x$ and $x'$, increases. (Note also that to avoid scaling issues with components of Euclidean distance, all variables are implicitly assumed to be standardized.) The practical implication of this Gaussian process approach is that for each finite collection, $X = (x_i : i = 1, .., N)$, of explanatory vectors in $\mathcal{X}$, the prior distribution of the associated random vector $y = y(X) = [y(x_i) : i = 1, .., N]$ is assumed to be *multinormal*:

$$y(X) \;\sim\; N[0_N,\; c_\omega(X,X)] \tag{3}$$

where $0_N$ denotes the $N$-vector of zeros and the covariance matrix, $c_\omega(X,X) \;=\; [c_\omega(x_i,x_j) : i,j = 1,..,N]$, is given by (2). Hence the entire process is defined by only the two parameters, $\omega = (v,\tau)$. While many extensions of this Gaussian process prior are possible that involve more parameters (as discussed further in the next section), our main objective is to show that with only a minimum number of parameters one can capture a wide range of complex nonlinear relationships.

Given this Gaussian process framework, the objective of GPR is to derive posterior predictions about unobserved $y$ values given observed values (data) at some subset of locations in $\mathcal{X}$. But here a new assumption is added, namely that observed values may themselves be subject to *measurement errors* that are independent of the actual process itself. Following Rasmussen and Williams (2006), we assume that for any realized value, $y(x)$, of the process at $x \in \mathcal{X}$, the associated *observed value*, $\tilde{y}(x)$, is a random variable of the form:

$$\tilde{y}(x) \;=\; y(x) \;+\; \varepsilon_x, \quad \varepsilon_x \underset{iid}{\sim} N(0,\sigma^2) \tag{4}$$

In this context, the relevant *prediction problem* for our purposes can be formulated as follows. Given observed data, $(\tilde{y},\tilde{X}) = \{(\tilde{y}_i,\tilde{x}_i),\; i = 1,..,n\}$, with $\tilde{y} = (\tilde{y}_i : i = 1,..,n)'$ and $\tilde{X} = (\tilde{x}_i : i = 1,..,n) \;\subset\; \mathcal{X}$, we seek to predict the unobserved value, $y(x)$, at $x \in \mathcal{X}$. To develop this prediction problem statistically, observe first from (3) and (4) that $\tilde{y}$ is multinormally distributed as

$$\tilde{y} \;\sim\; N\big[0_n\,,\; c_\omega(\tilde{X},\tilde{X}) + \sigma^2 I_n\big] \tag{5}$$

Hence, by a second application of (3), it follows that the prior distribution of $(y,\tilde{y})$ must be jointly multinormally distributed as (see e.g., expression (2.21) in Rasmussen and Williams 2006),

$$\begin{pmatrix} y \\ \tilde{y} \end{pmatrix} \;\sim\; N\left[\begin{pmatrix} 0 \\ 0_n \end{pmatrix},\; \begin{pmatrix} c_\omega(x,x) & c_\omega(x,\tilde{X}) \\ c_\omega(\tilde{X},x) & c_\omega(\tilde{X},\tilde{X}) + \sigma^2 I_n \end{pmatrix}\right] \tag{6}$$

Thus, by standard arguments (e.g., expression (A.6), p. 200 in Rasmussen and Williams 2006), one may conclude that the *conditional* distribution $y(x)$ given $(\tilde{y},\tilde{X})$, is of the form

$$y|x,\tilde{y},\tilde{X} \;\sim\; N\big[E(y|x,\tilde{y},\tilde{X}), \mathrm{var}(y|x,\tilde{y},\tilde{X})\big] \tag{7}$$

where by definition,

$$E(y|x,\tilde{y},\tilde{X}) \;=\; c_\omega(x,\tilde{X})\,[c_\omega(\tilde{X},\tilde{X}) + \sigma^2 I_n]^{-1}\,\tilde{y}\,, \text{ and} \tag{8}$$

$$\mathrm{var}(y|x,\tilde{y},\tilde{X}) \;=\; c_\omega(x,x) \;-\; c_\omega(x,\tilde{X})\,[c_\omega(\tilde{X},\tilde{X}) + \sigma^2 I_n]^{-1} c_\omega(\tilde{X},x) \tag{9}$$

This is usually referred to as the *predictive distribution* of $y(x)$ given observations, $(\tilde{y},\tilde{X})$. From a spatial modeling perspective, this predictive distribution is closely related to the method of geostatistical kriging (as discussed further in a more detailed version of this article, Dearmon and Smith (2014).

Up to this point, we have implicitly treated the parameters $(v, \tau, \sigma^2)$ as given. But in fact they are unknown quantities to be determined. Given the distributional assumptions above, one could employ empirical Bayesian estimation methods (as for example in Shi and Choi 2011, Section 3.1). But for our present purposes, it is most useful to adopt a *full Bayesian* approach in which all parameters are treated as random variables. This approach allows both parameter estimation and variable selection to be carried out simultaneously. In particular, the standard MCMC methods for Bayesian estimation allow model averaging methods to be used for both variable selection and parameter estimation. For purposes of this article, we adopt the approach developed in Chen and Wang (2010).[5]

First, to complete the full Bayesian specification of the model, we must postulate prior distributions for the vector of parameters,

$$\theta = (\omega, \sigma^2) \;=\; (v, \tau, \sigma^2) \;=\; (\theta_1, \theta_2, \theta_3) \tag{10}$$

Since these parameters are all required to be positive, we follow Chen and Wang (2010) (see also Williams and Rasmussen 1996) by postulating that they are independently log normally distributed with reasonably diffuse priors, and in particular that

$$\begin{aligned} \ln(\theta_1) &\sim N(-3, 9)\,, \\ \ln(\theta_2) &\sim N(3, 9)\,, \\ \ln(\theta_3) &\sim N(-3, 9) \end{aligned} \tag{11}$$

The prior for the length scale, $\theta_2$, is specified with a much larger mean. This prior predisposes the explanatory variables to be considered irrelevant. However, it is well know that so long as these prior distributions are independent and reasonably diffuse, their exact form will have little effect on the results. So the choices in (11) are largely a matter of convenience. If we now let $p(z)$ denote a generic probability density for any random vector, $z$, then for $z = \theta$, the full (*hyper*)*prior distribution* of $\theta$ can be written as

$$p(\theta) \;=\; \prod_{i=1}^{3} p(\theta_i) \tag{12}$$

where each of the marginals, $p(\theta_i)$, is a log normal density as in (11). Similarly, if we now let $z = \tilde{y}$, then the conditional distribution of $\tilde{y} = \tilde{y}(\tilde{X})$ given $\theta = (\omega, \sigma^2)$ is seen to be precisely the multinormal distribution in (5). So if for notational simplicity, we let

$$K_\theta(\tilde{X}) \;=\; c_\omega(\tilde{X}, \tilde{X}) + \sigma^2 I_n \tag{13}$$

then the corresponding conditional density, $p(\tilde{y}|\tilde{X}, \theta)$, is of the form

$$p(\tilde{y}|\tilde{X}, \theta) \;=\; (2\pi)^{-n/2} \det[K_\theta(\tilde{X})]^{-1/2} \exp\left[-\frac{1}{2}\tilde{y}'\, K_\theta(\tilde{X})^{-1}\tilde{y}\right] \tag{14}$$

Finally, if we assume that $\theta$ does not depend on $\tilde{X}$, that is, that $p(\theta|\tilde{X}) = p(\theta)$, then the desired *posterior distribution* of $\theta$ given data $(\tilde{y}, \tilde{X})$ can be obtained from the standard identity

$$p(\theta|\tilde{y}, \tilde{X})p(\tilde{y}|\tilde{X}) = p(\theta, \tilde{y}|\tilde{X}) = p(\tilde{y}|\tilde{X}, \theta)p(\theta|\tilde{X}) \;= p(\tilde{y}|\tilde{X}, \theta)p(\theta) \tag{15}$$

by noting that since $p(\tilde{y}|\tilde{X})$ does not involve $\theta$, we must have

$$p(\theta|\tilde{y},\tilde{X}) \;\propto\; p(\tilde{y}|\tilde{X},\theta)p(\theta) \tag{16}$$

A this point, one could in principle apply MCMC methods to estimate the posterior distribution of $\theta$ as well as posterior distributions of predictions, $y(x)$, in (7). But our goal is to combine such estimates with variable selection.

**Model and variable selection in GPR**

The above formulation of GPR has implicitly assumed that all explanatory variables, $x = (x_1, .., x_k)$, are relevant for describing variations in the response variable, $y$. But in most practical situations (such as our economic growth application below), it is important to be able to gauge which of these variables are most relevant. This is readily accomplished in standard regression settings where mean predictions are modeled as explicit functions of $x$, and hence where variable relevance can usually be tested directly in terms of associated parameters (such as in the standard linear specification, $E(y|x) = \beta_0 + \Sigma_{j=1}^{k}\beta_j x_j$). Even in the present GPR setting, there are a number of parametric approaches that have been proposed. The most popular of these is designated as ARD [see, e.g., MacKay (1995, 1998) and Neal (1996) together with the discussions in Rasmussen and Williams (2006, Section 5.1) and Shi and Choi (2011, Section 4.3.1)]. This method proceeds by the extending covariance model in (2) to include individual $\tau$ parameters for each variable,

$$c_\omega(x,x') \;=\; v\,\exp\left[-\frac{1}{2}\sum\nolimits_{j=1}^{k}\frac{(x_j - x'_j)^2}{\tau_j^2}\right] \tag{17}$$

where in this case, $\theta = (\omega, \sigma^2) = (v,\, \tau_1, .., \tau_k, \sigma^2)$. Here, it should be clear that for sufficiently large values of $\tau_j$ the variable $x_j$ will have little influence on covariance and hence on $y$ predictions. Hence the usual ARD procedure is to standardize all variables for comparability, construct estimates, $\hat{\tau}_j$, of $\tau_j$ by (empirical Bayes) maximum likelihood, and then determine some threshold value, $\tau_0$, for $\hat{\tau}_j$ above which $x_j$ is deemed to be irrelevant for prediction.

*BMA approach*

In contrast to this variable-selection procedure using extended parameterizations of the covariance kernel, our present approach essentially parameterizes "variable selection" itself. In particular, if we denote the presence or absence of each variable $x_j$ in a given model by the indicator function, $\delta_j$, with $\delta_j = 1$ if $x_j$ is present and $\delta_j = 0$ otherwise, then each model specification is defined by the values of the *model vector*, $\delta = (\delta_1, .., \delta_k)$. Here we omit the "null model," $\delta = 0_k$, and designate the set of possible values for $\delta$ as the *model space*, $\Delta = \{0,1\}^k - 0_k$. (This model-space approach to variable selection has a long history in Bayesian analysis, going back at least at to the work of George and McCulloch (1993) in hierarchical Bayesian regression.) With these definitions, one can now extend the set of model parameters, $\theta$, to include this model vector, $(\theta, \delta)$, and proceed to develop an appropriate prior distribution for $\delta$ on $\Delta$. In the present case, since the parameter vector, $\theta = (v, \tau, \sigma^2)$, is seen from (2) and (4) to be functionally independent of the choice of explanatory variables used (namely, $\delta$), we can assume that the priors on $\theta$ and $\delta$ are *statistically independent*.[6]

To construct a prior distribution for $\delta$, we first decompose this distribution as follows. If the *size* of each model, $\delta = (\delta_1, .., \delta_k)$, is designated by $q = s(\delta) = \sum_{j=1}^{k}\delta_j$, then by definition each prior, $p(\delta)$, for $\delta$ can be written as

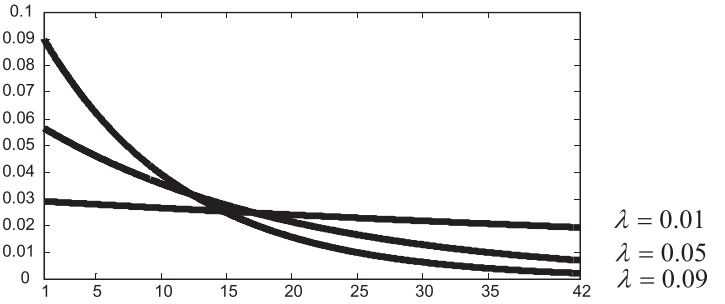


**Figure 1.** Selected values.

$$p(\delta) \;=\; p[\delta, s(\delta)] \;=\; p(\delta, q) \;=\; p(\delta|q)p(q) \tag{18}$$

This decomposition is motivated by the fact that the size of each model is itself an important feature. Indeed, all else being equal, smaller models are surely preferable to larger models (Occam's razor). So, it is reasonable to introduce some prior preference for smaller models. Following Chen and Wang, we employ a truncated geometric distribution for $q$ given by

$$p(q) \;=\; \frac{\lambda\,(1-\lambda)^{q-1}}{1-(1-\lambda)^{k}}, \quad q = 1, .., k \tag{19}$$

where $\lambda \in (0, 1)$. This family of distributions always places more weight on smaller values of $q$, as is seen in Fig. 1 below for selected values of $\lambda$ with $k = 42$. While Chen and Wang suggest that the (hyper)parameter, $\lambda$, be chosen by "tuning" the model (say with cross validation), we simply selected the prior value, $\lambda = 0.01$, which only slightly favors smaller values of $q$, as seen in the figure.

To complete the specification of $p(\delta)$ we assume that $p(\delta|q)$ is *uniform* on its domain. In particular, if for each $q = 1, .., k$ we let $\Delta_q = \{\delta \in \Delta : s(\delta) = q\}$ denote all models of size $q$, then the definition of $p(\delta)$ in (18) can be completed by setting

$$p(\delta|q) \;=\; \frac{1}{|\Delta_q|} \;=\; \frac{q!\,(k-q)!}{k!}\,; \quad \delta \in \Delta_q,\; q = 1, .., k \tag{20}$$

With this prior, we can now extend the posterior distribution in (16) to

$$p(\theta, \delta|\tilde{y}, \tilde{X}) \;\propto\; p(\tilde{y}|\tilde{X}, \theta, \delta)p(\theta, \delta) \;=\; p(\tilde{y}|\tilde{X}, \theta, \delta)p(\theta)p(\delta) \tag{21}$$

Following Chen and Wang, this joint posterior is estimated by Gibbs sampling using the conditional distributions,

$$p(\theta|\delta, \tilde{y}, \tilde{X}) \;\propto\; p(\tilde{y}|\tilde{X}, \theta, \delta)p(\theta), \text{ and} \tag{22}$$

$$p(\delta|\theta, \tilde{y}, \tilde{X}) \;\propto\; p(\tilde{y}|\tilde{X}, \theta, \delta)p(\delta) \tag{23}$$

We now consider each of these Gibbs steps in turn.

*Sampling the* $\theta$ *Posterior.* If for each component, $\theta_i$, of $\theta = (\theta_1, \theta_2, \theta_3)$ we now let $\theta_{-i}$ denote the vector of all other components, then (12) allows us to write the conditional distributions for these components as

$$p(\theta_i|\theta_{-i},\delta,\tilde{y},\tilde{X}) = \frac{p(\theta|\delta,\tilde{y},\tilde{X})}{p(\theta_{-i}|\delta,\tilde{y},\tilde{X})} \propto p(\theta|\delta,\tilde{y},\tilde{X}) \;\propto\; p(\tilde{y}|\theta,\delta,\tilde{X})p(\theta_i)p(\theta_{-i})$$
$$\propto\; p(\tilde{y}|\theta,\delta,\tilde{X})p(\theta_i),\; i=1,2,3 \tag{24}$$

Using these conditional distributions, one can in principle apply Gibbs sampling to approximate samples from the posterior in (22). But such samples are notoriously autocorrelated and cannot be treated as independent. This means that (in addition to the initial "burn in" samples) only a small fraction of these Gibbs samples can actually be used for analysis. With this in mind, we follow Chen and Wang by adopting an alternative approach designated as *Hamiltonian Monte Carlo* (HMC) (first introduced by Duane et al. (1987) and originally designated as "Hybrid Monte Carlo"). This approach not only requires a much smaller set of burn-in samples to reach the desired steady-state distribution (22), but can also be tuned to avoid autocorrelation problems almost entirely. The key idea (as developed in the lucid paper by Neal 2010) is to treat $\theta = (\theta_1,..,\theta_k)$ as the set of "position" variables in a discrete stochastic version of a $k$-dimensional Hamiltonian dynamical system with corresponding "momentum" variables, $\rho = (\rho_1,..,\rho_k)$. Such HMC processes can be tuned to converge to the desired steady-state distribution (22), while at the same time allowing extra "degrees of freedom" provided by the momentum variables, $\rho$. In particular, Neal (2010) shows how these momentum variables can be made to produce successive samples with "wider spacing" that tend to reduce autocorrelation effects.

*Sampling the* $\delta$ *Posterior.* In sampling from the posterior distribution of the model vector, $\delta$, we again follow Chen and Wang by employing a *Metropolis-Hastings* (M-H) algorithm with *birth-death* transition probabilities (see also Denison, Mallick, and Smith 1998). Since our method differs slightly from that of Chen and Wang, it is convenient to develop this procedure in more detail. The objective is to construct a Markov chain that converges to the distribution, $p(\delta|\theta,\tilde{y},\tilde{X})$, in (23). The basic "birth-death" idea is to allow only Markov transitions that add or subtract at most one variable from the current model. So if $\delta^q = (\delta_1^q,..,\delta_k^q)$ denotes a generic model of size $q$, then the possible "births" consist of those models in $\Delta_{q+1}$ that differ from $\delta^q$ by only one component, that is,

$$\Delta_{q+1}(\delta^q) \;=\; \Big\{\delta^{q+1} \in \Delta_{q+1} :\; \textstyle\sum_{i=1}^{k} |\delta_i^{q+1} - \delta_i^q| \;=1\Big\}, \quad q=1,..,k-1 \tag{25}$$

(where $\Delta_{q+1}(\delta^q) = \varnothing$ for $q = k$). Similarly, the possible "deaths" consist of those models in $\Delta_{q-1}$ that differ from $\delta^q$ in only one component, that is,

$$\Delta_{q-1}(\delta^q) \;=\; \Big\{\delta^{q-1} \in \Delta_{q-1} :\; \textstyle\sum_{i=1}^{k} |\delta_i^{q} - \delta_i^{q-1}| \;=1\Big\}, \quad q=2,..,k \tag{26}$$

(where $\Delta_{q-1}(\delta^q) = \varnothing$ for $q = 1$). With these definitions, the set of possible transitions, $\Delta(\delta^q)$, from each model, $\delta^q$, is of the form

$$\Delta(\delta^q) \;=\; \{\delta^q\} \cup \Delta_{q+1}(\delta^q) \cup \Delta_{q-1}(\delta^q), \quad q=1,..,k \tag{27}$$

If $T$ denotes the transition matrix for the desired Markov chain, and if we let $T(\delta|\delta^q)$ denote the corresponding *transition probability* from model $\delta^q$ to model $\delta \in \Delta(\delta^q)$, then by the general M-H algorithm, these transition probabilities are decomposed into the product of a

*proposal probability*, $p_r(\delta|\delta^q)$, and an *acceptance probability*, $p_a(\delta|\delta^q)$, for each $\delta \in \Delta(\delta^q) - \{\delta^q\}$ as

$$T(\delta|\delta^q) \;=\; p_r(\delta|\delta^q)\, p_a(\delta|\delta^q), \tag{28}$$

so that the "no transition" case is given by,

$$T(\delta^q|\delta^q) \;=\; 1 \;-\; \sum\nolimits_{\delta \in \Delta(\delta^q) - \{\delta^q\}} T(\delta|\delta^q) \tag{29}$$

In our case, the proposal probabilities are based on proposed "births" or "deaths." If we let $b$ denote a proposed *birth event* and $d$ a proposed *death event*, then by assuming these events are equally likely whenever both are possible, the appropriate *birth-death probability distribution*, $\pi(\cdot\,|\delta^q)$, can be defined as,

$$\pi(b|\delta^q) \;=\; \begin{cases} 1, & q = 1 \\ \dfrac{1}{2}, & 1 < q < k \\ 0, & q = k \end{cases} \text{, and} \tag{30}$$

$$\pi(d|\delta^q) \;=\; \begin{cases} 0, & q = 1 \\ \dfrac{1}{2}, & 1 < q < k \\ 1, & q = k \end{cases} \tag{31}$$

so that by definition, $\pi(b|\delta^q) + \pi(d|\delta^q) = 1$ for all $q = 1, .., k$. Given this birth-death process (which can be equivalently viewed as a random walk on $[1, .., k]$ with "reflecting barriers"), we next define *conditional proposal probabilities* given birth or death events. First, if $p_r(\delta|b, \delta^q)$ denotes the conditional probability of proposal, $\delta \in \Delta_{q+1}(\delta^q)$, given a *birth* event, $b$, and if all such proposals are taken to be equally likely, then since there are only $k - q$ ways of switching a "0" to "1" in $\delta^q$, it follows that

$$p_r(\delta|b, \delta^q) \;=\; \frac{1}{|\Delta_{q+1}(\delta^q)|} \;=\; \frac{1}{k-q}, \qquad \delta \in \Delta_{q+1}(\delta^q), \;\; q < k \tag{32}$$

Similarly, if $p_r(\delta|d, \delta^q)$ denotes the conditional probability of proposal, $\delta \in \Delta_{q-1}(\delta^q)$ given a *death* event, $d$, and if all such proposals are again taken to be equally likely, then since there are only $q$ ways of switching a "1" to "0" in $\delta^q$, it also follows that

$$p_r(\delta|d, \delta^q) \;=\; \frac{1}{|\Delta_{q-1}(\delta^q)|} \;=\; \frac{1}{q}, \qquad \delta \in \Delta_{q-1}(\delta^q), \;\; q > 1 \tag{33}$$

With these conventions, the desired proposal distribution in our case is given by

$$p_r(\delta|\delta^q) \;=\; \begin{cases} \pi(b|\delta^q) p_r(\delta|b, \delta^q), & \delta \in \Delta_{q+1}(\delta^q) \\ \pi(d|\delta^q) p_r(\delta|d, \delta^q), & \delta \in \Delta_{q-1}(\delta^q) \end{cases} \tag{34}$$

Finally, to ensure convergence to the posterior distribution, $p(\delta|\theta, \tilde{y}, \tilde{X})$, the desired *acceptance probability distribution*, $p_a(\cdot|\delta^q)$, for this M-H algorithm must necessarily be of the form

$$p_a(\delta|\delta^q) \;=\; \begin{cases} \min\{1,\, r(\delta,\delta^q)\}\;,\;\; \delta \in \Delta_{q+1}(\delta^q) \cup\, \Delta_{q-1}(\delta^q) \\ 1 - \sum_{\delta \in \Delta(\delta^q) - \{\delta^q\}} p_a(\delta|\delta^q),\;\; \delta = \delta^q \end{cases} \tag{35}$$

where the appropriate *acceptance ratio*, $r(\delta,\delta^q)$, is given by

$$r(\delta,\delta^q) \;=\; \frac{p(\delta|\theta,\tilde{y},\tilde{X})}{p(\delta^q|\theta,\tilde{y},\tilde{X})} \cdot \frac{p_r(\delta^q|\delta)}{p_r(\delta|\delta^q)}, \quad \delta \in \Delta_{q+1}(\delta^q) \cup\, \Delta_{q-1}(\delta^q) \tag{36}$$

As is shown in Supporting Information Appendix these ratios can be given the following operational form, where $p(q)$ denotes the truncated geometric distribution in (19)

$$r(\delta,\delta^q) \;=\; \begin{cases} \dfrac{p(\tilde{y}|\delta,\theta,\tilde{X})}{p(\tilde{y}|\delta^q,\theta,\tilde{X})} \cdot \dfrac{p(q+1)}{p(q)} \cdot \dfrac{\pi(d|\delta)}{\pi(b|\delta^q)},\;\; \delta \in \Delta_{q+1}(\delta^q) \\ \dfrac{p(\tilde{y}|\delta,\theta,\tilde{X})}{p(\tilde{y}|\delta^q,\theta,\tilde{X})} \cdot \dfrac{p(q-1)}{p(q)} \cdot \dfrac{\pi(b|\delta)}{\pi(d|\delta^q)},\;\; \delta \in \Delta_{q-1}(\delta^q) \end{cases} \tag{37}$$

*Gibbs Sampling*. The basic Gibbs sampling procedure outlined above was programmed in MATLAB (and is described in more detail in the appendix of Dearmon and Smith (2014). The M-H algorithm for sampling model vectors, $\delta$, forms the outer loop of this procedure, and the HMC procedure for sampling parameter vectors, $\theta$, forms the inner loop. This structure allows more efficient sampling, depending on whether new model vectors are chosen or not. Following an initial burn-in phase, a post burn-in sequence, $[(\delta_i,\theta_i) : i = 1,..,N]$, is obtained for estimating all additional properties of this Gaussian process model, as detailed below.

*Model probabilities and variable-inclusion probabilities*

With regard to the general problem of model selection, one of the chief advantages of this model-space approach is that it yields meaningful posterior probabilities for each candidate model vector, $\delta$, given the observed data $(\tilde{y},\tilde{X})$. In particular, these *model probabilities* are simply the marginal probabilities,

$$p(\delta|\tilde{y},\tilde{X}) \;=\; \int_\theta p(\delta,\theta|\tilde{y},\tilde{X})\, d\theta \tag{38}$$

For estimation purposes, it is more convenient to write these probabilities as conditional expectations over the entire space of parameter pairs, $(\delta,\theta)$. In particular, if for each model $\delta_\alpha \in \Delta$ we let the indicator function, $I_\alpha(\delta,\theta)$, be defined by

$$I_\alpha(\delta,\theta) \;=\; \begin{cases} 1, & \delta = \delta_\alpha \\ 0, & \delta \neq \delta_\alpha \end{cases} \tag{39}$$

then (38) can be equivalently written as a general integral of the form

$$p(\delta_\alpha|\tilde{y},\tilde{X}) \;=\; \int_{(\delta,\theta)} I_\alpha(\delta,\theta)\, p(\delta,\theta|\tilde{y},\tilde{X})(d\delta \times d\theta) \;=\; E_{(\delta,\theta)}[I_\alpha(\delta,\theta)] \tag{40}$$

But since these are the steady-state probabilities for the (irreducible) Markov process with realizations, $[(\delta_i, \theta_i) : i = 1, .., N]$, the ergodic properties of such processes are well known to imply that the *sample average*,

$$\hat{p}(\delta_\alpha|\tilde{y}, \tilde{X}) \;=\; \frac{1}{N}\sum\nolimits_{i=1}^{N} I_\alpha(\delta_i, \theta_i) \tag{41}$$

converges to $p(\delta_\alpha|\tilde{y}, \tilde{X})$ with probability one. Moreover, since the number of occurrences of $\delta_\alpha$ in the sample sequence, $[(\delta_i, \theta_i) : i = 1, .., N]$, is given by,

$$N(\delta_\alpha) \;=\; \sum\nolimits_{i=1}^{N} I_\alpha(\delta_i, \theta_i) \tag{42}$$

it follows from (41) that for any model, $\delta \in \Delta$, this sample average is simply the fraction of $\delta$ occurrences, that is,

$$\hat{p}(\delta|\tilde{y}, \tilde{X}) \;=\; \frac{N(\delta)}{N} \tag{43}$$

Note also that (43) yields an estimate, $\hat{\delta}$, of the *most likely model* based on observations, $(\tilde{y}, \tilde{X})$, namely

$$\hat{\delta} \;=\; \arg\max\nolimits_{\delta\in\Delta} \hat{p}(\delta|\tilde{y}, \tilde{X}) \;=\; \arg\max\nolimits_{\delta\in\Delta} \frac{N(\delta)}{N} \tag{44}$$

In this context, one might be tempted to identify the "most relevant" explanatory variables in $x = (x_1, .., x_j, .., x_k)$ to be simply those appearing in this most likely model. But like the ARD procedure mentioned above, this method provides no probabilistic measure of "relevance" for each variable separately. However, in a manner similar to posterior likelihoods of models, we can also define posterior likelihoods of individual variables as follows. If we denote the class of models containing variable $j$ by $\Delta_j = \{\delta \in \Delta : \delta_j = 1\}$, then in terms of model probabilities, it follows that the probable membership of variable $j$ in such candidate models must be given by

$$p(\delta_j = 1|\tilde{y}, \tilde{X}) \;=\; \sum\nolimits_{\delta\in\Delta_j} p(\delta|\tilde{y}, \tilde{X}) \tag{45}$$

Moreover, we see from (41) that a consistent estimator of this *inclusion probability* for each variable $j$ is given by

$$\hat{p}(\delta_j = 1|\tilde{y}, \tilde{X}) \;=\; \sum\nolimits_{\delta\in\Delta_j} \hat{p}(\delta|\tilde{y}, \tilde{X}) \;=\; \sum\nolimits_{\delta\in\Delta_j} \frac{N(\delta)}{N} \tag{46}$$

Finally, since

$$N_j \;=\; \sum\nolimits_{\delta\in\Delta_j} N(\delta) \tag{47}$$

is by definition the number of occurrences of variable $j$ in the models of sample sequence, $[(\delta_i, \theta_i) : i = 1, .., N]$, it follows as a parallel to (43) that this estimated inclusion probability is simply the fraction of these occurrences,

$$\hat{p}(\delta_j = 1|\tilde{y}, \tilde{X}) \;=\; \frac{N_j}{N} \tag{48}$$

These inclusion probabilities provide a natural measure of relevance for each variable which (unlike *P*-values) is *larger* for more relevant variables. For example, if the estimated inclusion probability for a given variable, $j$, is 0.95, then $j$ must appear in 95% of the (post burn-in) models "accepted" by the Metropolis-Hastings procedure above. So while there is no formal "null hypothesis" being tested, this inclusion probability does indeed provide compelling evidence for the relevance of variable $j$ based on observations, $(\tilde{y}, \tilde{X})$.

*Prediction and marginal effects using BMA*

One key difference between inclusion probabilities and standard tests of hypotheses for regression coefficients is that inclusion probabilities yield no direct information about whether the contribution of a given explanatory variable tends to be positive or negative. In fact, when relations among variables are highly nonseparable (as in our examples below), both the magnitude and direction of such contributions may exhibit substantial local variation. In view of these possibilities, it is more appropriate to consider the *local* contributions of each component of $x = (x_1, .., x_k)$ to predicted values of the response variable, $y(x)$. With this objective in mind, we first employ the MCMC results above to develop posterior mean predictions of $y(x)$ given $(\tilde{y}, \tilde{X})$ that parallel expression (8) above.

*BMA Predictions.* To obtain posterior mean predictions, one could in principle apply the post burn-in sequence, $[(\delta_i, \theta_i) : i = 1, .., N]$, to estimate maximum a posteriori (MAP) values, $\hat{\theta} = (\hat{\omega}, \hat{\sigma}^2)$, of the parameters together with the most likely model, $\hat{\delta}$, in (44) and use this pair $(\hat{\delta}, \hat{\theta})$ to obtain a posterior version of the mean predictions in (8). In particular, if for any data point, $x = (x_1, .., x_k) \in \mathcal{X}$, we now denote the relevant data for each model, $\delta \in \Delta$, by $x(\delta) \;=\; (x_j : \delta_j = 1)$, and similarly, let $\tilde{X}(\delta) \;=\; [\tilde{x}_1(\delta), .., \tilde{x}_n(\delta)]$, then using (8) together with (13), the *MAP prediction* of $y$ given $x(\hat{\delta})$ together with data, $[\tilde{y}, \tilde{X}(\hat{\delta})]$, can be obtained as,

$$E[y|x(\hat{\delta}), \tilde{y}, \tilde{X}(\hat{\delta})] \;=\; c_{\hat{\omega}}[x(\hat{\delta}), \tilde{X}(\hat{\delta})] \left\{K_{\hat{\theta}}[\tilde{X}(\hat{\delta})]\right\}^{-1} \tilde{y} \tag{49}$$

However, as is widely recognized, there is often more information in the underlying MCMC sequence $[(\delta_i, \theta_i) : i = 1, .., N]$ than is provided by this single MAP estimate. In particular, by averaging the mean predictions generated by each of the sample pairs, $(\delta_i, \theta_i)$, the resulting "ensemble" prediction is generally considered to be more robust. This is in fact the essence of BMA.

So rather than using (49), we now construct BMA predictions of $y$ (as first proposed by Raftery, Madigan, and Hoeting 1997). To do so, recall first (from the "Gaussian process regression" Section) that the spatial location of any prediction may or may not be part of the candidate variables in $x$ (let alone the reduced variable set, $x(\delta)$, for any model, $\delta \in \Delta$). But for purposes of spatial prediction, it is useful to be explicit about the underlying set of *locations*, $l \in L$. For any given location, $l$, we now let $x_l = (x_{l1}, .., x_{lk}) \in \mathcal{X}$ denote the vector of candidate explanatory variables at $l$, and let $y_l$ denote the corresponding value of $y$ to be predicted at $l$. By replacing $(\hat{\delta}, \hat{\theta})$ in (49) with the pair, $(\delta_i, \theta_i)$, the corresponding mean prediction for $y_l$ given $(\delta_i, \theta_i)$ together with data $[\tilde{y}, \tilde{X}]$ is then given by:

$$E\left[y_l|x_l,\tilde{y},\tilde{X},\delta_i,\theta_i\right] \;=\; c_{\omega_i}[x_l(\delta_i),\tilde{X}(\delta_i)]\left\{K_{\theta_i}[\tilde{X}(\delta_i)]\right\}^{-1}\tilde{y},\quad i=1,..,N \tag{50}$$

In these terms, the corresponding *BMA prediction* of $y_l$ at location, $l \in L$, is given by

$$E(y_l|x_l,\tilde{y},\tilde{X}) \;=\frac{1}{N}\sum\nolimits_{i=1}^{N} E[y_l|x_l,\tilde{y},\tilde{X},\delta_i,\theta_i] \tag{51}$$

Note in particular that such mean predictions are equally well defined at *data points* $(\tilde{y}_j,\tilde{x}_j),\ j=1,..,n$, and are given by

$$\begin{aligned} E(\tilde{y}_j|\tilde{x}_j,\tilde{y},\tilde{X}) \;&=\; \frac{1}{N}\sum\nolimits_{i=1}^{N} E[\tilde{y}_j|\tilde{x}_j,\tilde{y},\tilde{X},\delta_i,\theta_i] \\ &=\; \frac{1}{N}\sum\nolimits_{i=1}^{N} c_{\omega_i}[\tilde{x}_j(\delta_i),\tilde{X}(\delta_i)]\left\{K_{\theta_i}[\tilde{X}(\delta_i)]\right\}^{-1}\tilde{y} \end{aligned} \tag{52}$$

*BMA Marginal Effects*. Here, we again adopt a BMA approach to local marginal effects at locations, $l \in L$, by first considering these effects for each mean prediction in (50), and then averaging such effects as in (51). Turning first to the mean predictions in (50) generated by a given pair, $(\delta_i,\theta_i)$, there are several issues that need to be addressed. First, there is the question of how to treat components of $x_l$ that are excluded from model, $\delta_i$. One approach is simply to ignore such cases by only calculating marginal effects for each explanatory variable, $x_{lj}$, in those models, $\delta_i$, with $\delta_{ij}=1$, and then averaging these effects. But for purposes of model averaging, it is more appropriate to simply treat marginal effects as being identically zero for excluded variables. (These two approaches are compared following expression (58) below.) The second question is how to calculate local marginal effects for included variables. For simplicity, we here treat all variables, $x_{lj}$, as continuous, and thus define their marginal effects to be simply the partial derivatives of mean predictions in (50) with respect to $x_{lj}$.

With these preliminaries, we now define the *marginal effect*, $ME_{lj}(\delta_i,\theta_i)$, of explanatory variable, $j$, in the $(\delta_i,\theta_i)$-prediction at location, $l$, to be:

$$ME_{lj}(\delta_i,\theta_i) \;=\; \begin{cases} \dfrac{\partial}{\partial x_j} E\left[y_l|x_l,\tilde{y},\tilde{X},\delta_i,\theta_i\right], & \delta_{ij}=1 \\ 0, & \delta_{ij}=0 \end{cases} \tag{53}$$

For the case of $\delta_{ij}=1$, we may use (50) to obtain the following more explicit form[7]:

$$\frac{\partial}{\partial x_j} E\left[y_l|x_l,\tilde{y},\tilde{X},\delta_i,\theta_i\right] \;=\; \left\{\frac{\partial}{\partial x_j} c_{\omega_i}[x_l(\delta_i),\tilde{X}(\delta_i)]\right\} K_{\theta_i}[\tilde{X}(\delta_i)]^{-1}\tilde{y} \tag{54}$$

Moreover, since partial derivatives of the squared exponential kernel in (2) are given by

$$\frac{\partial}{\partial x_j} c_\omega(x,\tilde{x}) \;=\; \frac{\partial}{\partial x_j}\left[v\,\exp(-\frac{1}{2\tau^2}||x-\tilde{x}||^2)\right] \;=\; c_\omega(x,\tilde{x})[-\frac{1}{\tau^2}(x_j-\tilde{x}_j)] \tag{55}$$

it follows by letting $x_l(\delta_i)=x_{li}$ and $\tilde{X}(\delta_i)=[\tilde{x}_{i1},..,\tilde{x}_{in}]$ that the bracketed expression in (54) can be given the following exact form

$$\begin{aligned}\frac{\partial}{\partial x_j} c_{\omega_i}[x_l(\delta_i), \tilde{X}(\delta_i)] &= [\frac{\partial}{\partial x_j} c_{\omega_i}(x_{li}, \tilde{x}_{i1}), ..., \frac{\partial}{\partial x_j} c_{\omega_i}(x_{li}, \tilde{x}_{in})] \\ &= -\frac{1}{\tau_i^2}\,[c_{\omega_i}(x_{li}, \tilde{x}_{i1})(x_{lij} - \tilde{x}_{i1j}), .., c_{\omega_i}(x_{li}, \tilde{x}_{in})(x_{lij} - \tilde{x}_{inj})]\end{aligned} \tag{56}$$

As in (51), the resulting BMA *marginal effect*, $ME_{lj}$, of explanatory variable, $j$, at location, $l$, is simply the average of the values in (53) as given by

$$ME_{lj} = \frac{1}{N} \sum\nolimits_{i=1}^{N} ME_{lj}(\delta_i, \theta_i) \tag{57}$$

Note finally that since all terms with $\delta_{ij} = 0$ are zero, and since the number of models, $\delta_i$, with $\delta_{ij} = 1$ is precisely $N_j$ in (47), this marginal effect can be equivalently written as

$$ME_{lj} = \frac{1}{N}\left[\sum\nolimits_{i\,:\,\delta_{ij}=1} ME_{lj}(\delta_i, \theta_i)\right] = \frac{N_j}{N}\left[\frac{1}{N_j}\sum\nolimits_{i\,:\,\delta_{ij}=1} ME_{lj}(\delta_i, \theta_i)\right] \tag{58}$$

The expression in brackets is precisely the BMA marginal effect that would have been obtained if only models involving variable $j$ were included in the averaging. Hence the present version simply "discounts" marginal effects by the inclusion probabilities in (48).

Given this formal development of GPR-BMA models, we turn now to a systematic comparison of this approach with the alternative approaches outlined in the Introduction, namely, the BMA versions of spatial regression models proposed by LeSage and Parent (2007).

## Simulation study

As mentioned in the Introduction, the simple simulation models developed here are designed specifically to focus on the role of functional nonseparabilities in comparing GPR-BMA with SAR-BMA and SEM-BMA. To do so, it is appropriate to begin in the next section with a brief description of these spatial regression models. This is followed by a specification of the simulation models to be used, together with comparative simulation results focusing on variable selection and marginal effect estimation.

### SAR-BMA and SEM-BMA models

Following LeSage and Parent (2007), the standard *SAR model* takes the form

$$y = \rho W y + \alpha\, \iota_n + X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n) \tag{59}$$

where $X = [x_i : i = 1, .., k]$ is an $n \times k$ matrix of explanatory variables (as in the "Gaussian process regression" Section above), and where $\iota_n = (1, .., 1)'$ is a unit vector representing the intercept term in the regression. The key new element here is the prior specification of an $n$-square weight matrix, $W$, which is taken to summarize all spatial relations between sample locations, $i = 1, .., n$. For purposes of analysis, the simultaneities among dependent values in $y$ are typically removed by employing the reduced form:

$$y = (I_n - \rho W)^{-1} (\alpha \iota_n + X\beta + \varepsilon), \quad \varepsilon \sim N(0, \sigma^2 I_n) \tag{60}$$

In terms of this same notation, the standard *SEM model* is given by the equation system,

$$y = \alpha \iota_n + X\beta + u, \quad u = \rho W u + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n) \tag{61}$$

where simultaneities among residual values in $u$ are similarly removed by employing the reduced form:

$$y = \alpha \iota_n + X\beta + (I_n - \rho W)^{-1} \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n) \tag{62}$$

Note that (unlike LeSage and Parent) we use the same symbol, $\rho$, for the spatial autocorrelation parameter in both models (60) and (62) to emphasize the similarity in parameter sets, $(\alpha, \beta, \sigma, \rho)$, between these models. Not surprisingly, this similarity simplifies extensions to a Bayesian framework, since one can often employ common prior distributions for parameters in both models. To facilitate BMA, LeSage and Parent follow many of the general conventions proposed by Fernandez, Ley, and Steel (2001b). First of all (in a manner similar to our $\delta$ vectors in the "BMA approach" Section above), if the relevant class of candidate models, $M$, is denoted by $\mathbb{M}$, then each model, $M \in \mathbb{M}$, is specified by a selection of variables (columns) from $X$, denoted here by $X_M$, with corresponding parameter vector, $\beta_M$. The parameters $\alpha$ and $\sigma$ together with the relevant spatial autocorrelation parameter, $\rho$, are estimated for each model, and are given standard noninformative priors. In particular a uniform prior on $[-1,1]$ is adopted for $\rho$ in all simulations below. Only the priors on $\beta_M$ for each model $M$ warrant further discussion, since they utilize data information from $X_M$. In particular, the prior on $\beta_M$ for SAR-BMA models is assumed to be normal with mean vector, 0, and covariance matrix given by $g(X'_M X_M)^{-1}$, where (following the recommendation of Fernandez, Ley, and Steel 2001a) the proportionality factor is given by $g = 1/\max\{n, k^2\}$, with $k$ denoting the number of candidate explanatory variables. As with our GPR-BMA model, all variables are here assumed to be *standardized*, both to be consistent with the zero prior mean assumption on $\beta_M$ and to avoid sensitivity to units in the associated covariance matrix. For SEM-BMA models, the prior on $\beta_M$ is given a similar form, with $X_M$ replaced by $(I_n - \lambda W)X_M$. In both cases, these covariances are motivated by standard maximum-likelihood estimates of $\beta_M$, and can thus be said to yield natural "empirical Bayes" priors for $\beta_M$.

Aside from the specification of priors, the other key difference between the implementation of SAR-BMA and SEM-BMA in LeSage and Parent 2007 and our implementation of GPR-BMA in the "BMA approach" Section above is the method of estimating both model probabilities and inclusion probabilities. Rather than appeal to asymptotic MCMC frequency approximations as we have done, LeSage and Parent follow the original approach of Fernandez, Ley, and Steel (2001a) by employing numerical integration to obtain direct approximations of the posterior marginal probabilities for each model. If we again let $(\tilde{y}, \tilde{X})$ denote the relevant set of observed data as in the "Model and variable selection in GPR" Section above, and let $p(M|\tilde{y}, \tilde{X})$ denote the posterior marginal probability of model $M$ given $(\tilde{y}, \tilde{X})$, then the corresponding *estimated model probabilities*, $\hat{p}(M|\tilde{y}, \tilde{X})$, are taken to be these numerical-integration approximations. If for each variable, $j$, we also let $\mathbb{M}_j$ denote the set of models, $M$, containing variable $j$, then as a parallel to expression (46) above, the relevant estimates of *inclusion probabilities* for each variable $j$ is given by

$$\hat{p}(j|\tilde{y},\tilde{X}) = \sum\nolimits_{M\in\mathbb{M}_j} \hat{p}(M|\tilde{y},\tilde{X}) \tag{63}$$

As verified by Fernandez, Ley, and Steel (2001b), both the frequency and numerical-integration approaches yield very similar results for sufficiently large MCMC sample sizes. But since the posterior marginal calculations should in principle be somewhat more accurate, they can be expected to give a slight "edge" to both SAR-BMA and SEM-BMA simulations (based on the MATLAB routines of LeSage) over our asymptotic frequency approach for GPR-BMA. This lends further weight to the marked superiority of GPR-BMA estimates as exhibited by the simulations below.

Finally, with respect to local marginal effects, it follows from (62) that such effects are constant across space for SEM and are given for all $ij$ by $\partial E(Y_i|X)/\partial x_{ij} \equiv \beta_j$. Thus, these effects in SEM-BMA are simply Bayesian model averages of the estimates, $\hat{\beta}_j$, over all models.[8] But as seen in (60), such effects are more complex for SAR. If $b^{ii}_\rho$ denotes the $i$ th diagonal element of the matrix, $B_\rho = (I_n - \rho W)^{-1}$, then it can be shown (LeSage and Pace 2009, Section 2.7.1) that $\partial E(Y_i|X)/\partial x_{ij} = \beta_j b^{ii}_\rho$, so that SAR-BMA estimates these effects as Bayesian model averages of $\hat{\beta}_j b^{ii}_{\hat{\rho}}$ over all models. However, it must be emphasized that since the multipliers, $b^{ii}_{\hat{\rho}}$, are positive in almost all cases of interest, the *signs* of these effects tend to be constant across space. In empirical settings where $\hat{\rho}$ values are relatively constant across models, the *relative magnitudes*, $E[\hat{\beta}_j b^{ii}_{\hat{\rho}}]/E[\hat{\beta}_k b^{ii}_{\hat{\rho}}] \approx \hat{\beta}_j/\hat{\beta}_k$, are also approximately constant and such "local" effects actually exhibit little qualitative variation over space (as is evidenced by our empirical application below).

**Simulated model comparisons**

We start with the following 3-variable instance of the SAR model in (60),

$$y \;=\; (I_n - \rho W)^{-1}\,[3\,\iota_n + \;x_1 + 4x_2 - 2\,x_3 + \;\varepsilon], \quad \varepsilon \sim N(0, \sigma^2 I_n) \tag{64}$$

and corresponding instance of the SEM model in (62),

$$y \;=\; 3\,\iota_n + \;x_1 + 4x_2 - 2\,x_3 + \;(I_n - \rho W)^{-1}\varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n) \tag{65}$$

where in both cases, $X = (x_1, x_2, x_3)$, $\alpha = 3$, and $\beta = (1, 4, -2)'$. In view of the linear separability of these specifications, we designate these benchmark models as the *separable models*. Our main interest will be in the behavior of estimators when the actual functional form is not separable. But before introducing such complexities, we first complete the parameter specification of the basic models in (64) and (65). For all simulations in this section, we set the autocorrelation parameter to $\rho = 0.5$ (to ensure a substantial degree of spatial autocorrelation), and choose a sample size, $n = 367$, that is sufficiently large to avoid small-sample effects. In particular, the weight matrix, $W$, used here is a queen-contiguity matrix for Philadelphia census tracts (normalized to have a maximum eigenvalue of one). Finally, the simulated values of $(x_1, x_2, x_3)$ are standardizations of independent samples drawn from $N(0, 1)$, and the residual standard deviation is set to be sufficiently small, $\sigma = 0.1$, to ensure that functional specifications of $(x_1, x_2, x_3)$ always dominate residual noise. In this setting, it is clear that both SAR and SAR-BMA should do very well in estimating model (64), and similarly that both SEM and SEM-BMA should do well for (65).

To introduce nonseparabilities into these models, we preserve all spatial autocorrelation specifications, but alter the functional form of $(x_1, x_2, x_3)$ as follows[9]:

$$y = (I_n - \rho W)^{-1}[3\,\iota_n + x_1 \cdot (4x_2 - 2x_3) + \varepsilon], \quad \varepsilon \sim N(0, \sigma^2 I_n) \tag{66}$$

$$y = 3\,\iota_n + x_1 \cdot (4x_2 - 2x_3) + (I_n - \rho W)^{-1}\varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n) \tag{67}$$

This seemingly "innocent" change serves to highlight the main objective of the present analysis. In particular, it should be clear that the effective sign of $x_1$ now depends on the sign of $4x_2 - 2x_3$, and similarly that the effective signs of both $x_2$ and $x_3$ depend on that of $x_1$. So, a key feature of these *nonseparable models* is that the direction of influence of each $x$-variable on $y$ depends on the values of other $x$-variables. This we refer to as a localized marginal effect (LME) which, in principle, can be quite different from the average marginal effect (AME). Thus, it should be clear that any attempt to approximate such nonseparabilities by appropriate choices of (constant) coefficients, $\beta$, in $X\beta$ is bound to fail. Even more important is the fact that such "compromise" approximations may often be so close to zero that the explanatory variables are rendered *statistically and quantitatively insignificant*. This is in fact a main conclusion of our simulation results.

But before presenting these results, it is important to observe that models (66) and (67) can of course be well estimated by simply extending the linear-in-parameters specifications in (64) and (65) to include first-order interaction effects. In particular, since the expression $x_1(4x_2 - 2x_3) = 4x_1x_2 - 2x_1x_3$ is an instance of the 6-parameter specification, $\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3$, standard estimates of models (60) and (62) with $X$ extended to $Z = (X, x_1x_2, x_1x_3, x_2x_3)$, can easily identify the two significant parameters, $\beta_4$ and $\beta_5$. More generally, such parametric specifications can in principle be extended to capture almost any degree of interaction complexity. But such heavily parameterized ("saturated") models are not only costly in terms of data requirements, they are also notoriously prone to over-fitting data. These points serve to underscore our emphasis on the ability of GPR-BMA to identify highly complex relations with remarkably few parameters. Finally, to gauge the effectiveness of each method in identifying the true explanatory variables, $(x_1, x_2, x_3)$, three irrelevant variables $(z_1, z_2, z_3)$, are also constructed as standardizations of independent samples from $N(0, 1)$, and added to each simulation.

**Simulation results**

Results are displayed in Table 1 as a series of two panels: SAR is the top panel, SEM the bottom panel. Each panel is divided into two general sections corresponding to the separable and nonseparable cases, respectively. The first two columns of each panel display the relevant true values for each model, and the remaining columns contain the comparative results to be discussed below.

But before doing so, it is important to re-emphasize that we have made certain assumptions that are particularly favorable to SAR-BMA and SEM-BMA. Of these assumptions, the most significant are that the data generating processes are precisely of the SAR and SEM forms, and in particular, that the researcher has chosen the correct spatial weights matrix in each case. In contrast, none of these assumptions are made by GPR-BMA. Indeed, except for the choice of a generic kernel function such as (2), GPR-BMA is essentially driven by the data itself.[10] As we will show, the results of GPR-BMA are thus more robust in situations where the researcher faces both specification and model uncertainty.

With respect to variable inclusion probabilities (VIP) for separable cases, GPR-BMA performs as well or better than the traditional methods in essentially every case (the only

**Table 1.** Simulation Results: Top Panel—SAR, Bottom Panel—SEM

| | Separable | | | | | | | | Nonseparable | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | True | | VIP | | AME-Values | | LME- RMSE | | True | | VIP | | AME-Values | | LME- RMSE | | Signs- % Correct | |
| | VIP | AME | SAR-BMA | GPR-BMA | SAR-BMA | GPR-BMA | SAR-BMA | GPR-BMA | V.I.P. | AME | SAR-BMA | GPR-BMA | SAR-BMA | GPR-BMA | SAR-BMA | GPR-BMA | SAR-BMA | GPR-BMA |
| x1 | 1.000 | 1.039 | **1.000** | **1.000** | **1.036** | 1.101 | **0.003** | 0.121 | 1.000 | 0.000 | 0.663 | **1.000** | 0.401 | **0.100** | 4.396 | **0.345** | 51% | **98%** |
| x2 | 1.000 | 4.157 | **1.000** | **1.000** | **4.159** | 4.118 | **0.001** | 0.146 | 1.000 | 0.000 | 0.068 | **1.000** | **−0.012** | 0.029 | 4.015 | **0.381** | 52% | **99%** |
| x3 | 1.000 | −2.079 | **1.000** | **1.000** | **−2.068** | −2.038 | **0.011** | 0.059 | 1.000 | 0.000 | 0.050 | **1.000** | **0.001** | 0.070 | 2.008 | **0.240** | 52% | **98%** |
| z1 | 0.000 | 0.000 | **0.050** | 0.153 | **0.000** | −0.015 | **0.000** | 0.020 | 0.000 | 0.000 | 0.068 | **0.000** | 0.013 | **0.000** | 0.013 | **0.000** | 0% | **100%** |
| z2 | 0.000 | 0.000 | 0.050 | **0.018** | **0.000** | 0.000 | **0.000** | 0.001 | 0.000 | 0.000 | 0.073 | **0.000** | −0.015 | **0.000** | 0.015 | **0.000** | 0% | **100%** |
| z3 | 0.000 | 0.000 | 0.050 | **0.000** | **0.000** | 0.000 | **0.000** | **0.000** | 0.000 | 0.000 | 0.050 | **0.000** | −0.001 | **0.000** | 0.001 | **0.000** | 0% | **100%** |
| | VIP | AME | SEM-BMA | GPR-BMA | SEM-BMA | GPR-BMA | SEM-BMA | GPR-BMA | VIP | AME | SEM-BMA | GPR-BMA | SEM-BMA | GPR-BMA | SEM-BMA | GPR-BMA | SEM-BMA | GPR-BMA |
| x1 | 1.000 | 1.000 | **1.000** | **1.000** | **1.001** | 0.998 | **0.001** | 0.010 | 1.000 | 0.000 | 0.071 | **1.000** | 0.014 | **−0.001** | 4.447 | **0.085** | 51% | **100%** |
| x2 | 1.000 | 4.000 | **1.000** | **1.000** | 4.005 | **4.000** | **0.005** | 0.012 | 1.000 | 0.000 | 0.055 | **1.000** | −0.006 | **−0.002** | 3.995 | **0.079** | 49% | **100%** |
| x3 | 1.000 | −2.000 | **1.000** | **1.000** | **−1.999** | −1.996 | **0.001** | 0.011 | 1.000 | 0.000 | 0.204 | **1.000** | 0.081 | **0.000** | 1.999 | **0.058** | 49% | **100%** |
| z1 | 0.000 | 0.000 | 0.050 | **0.000** | **0.000** | **0.000** | **0.000** | **0.000** | 0.000 | 0.000 | 0.068 | **0.000** | −0.013 | **0.000** | 0.013 | **0.000** | 0% | **100%** |
| z2 | 0.000 | 0.000 | 0.050 | **0.000** | **0.000** | **0.000** | **0.000** | **0.000** | 0.000 | 0.000 | 0.265 | **0.000** | −0.116 | **0.000** | 0.116 | **0.000** | 0% | **100%** |
| z3 | 0.000 | 0.000 | 0.071 | **0.000** | 0.001 | **0.000** | 0.001 | **0.000** | 0.000 | 0.000 | 0.081 | **0.000** | 0.018 | **0.000** | 0.018 | **0.000** | 0% | **100%** |

Boldface indicates best performance.

exception being the irrelevant variable, $z_1$, with inclusion probability slightly higher than for SAR-BMA, but still very small). Turning to the nonseparable cases, we see a much more dramatic difference in performance. Here, (as predicted) both SAR-BMA and SEM-BMA are seen to have great difficulty identifying any of the statistically relevant explanatory variables (where only the inclusion probability for variable, $x_1$, under SAR-BMA is noticeably better than those of the irrelevant variables). In contrast, the performance of GPR-BMA is quite striking, with essentially perfect identification of relevant variables and exclusion of irrelevant variables.

Turning next to marginal effects, we first calculated SAR and SEM coefficient estimates for each model identified by the BMA routine (as in footnote 8 above). These estimates were in turn used to calculate LMEs for each model (which are by definition constant for SEM, and thus the same as AMEs). For SAR, these localized effects were used together with model inclusion probabilities to calculate AMEs. While such effects are not strictly part of SAR-BMA and SEM-BMA as formulated by LeSage and Parent, they appear to provide a reasonable basis of comparison with GPR-BMA.

For the separable cases, the AMEs for GPR-BMA are seen to be remarkably comparable with those of SAR-BMA and SEM-BMA, even though none of the true model structure is assumed in GPR-BMA. To compare AME for nonseparable cases, it should first be noted that all true AMEs for the simulated model are necessarily close to zero (Column 10, both panels) since marginals are linear in the standardized $X$ variables, with sample means close to zero. Moreover, since the uniformly low VIP for both SAR-BMA and SEM-BMA tend to shrink their marginals toward zero,[11] our simulation choice actually tends to favor these models in terms of AME. So, it is remarkable that GPR-BMA still does better than either in most cases (Column 14, both panels).

Turning next to LMEs, we here measure performance in terms of overall goodness of fit [using Root Mean Squared Error (RMSE)] at the individual observation level. For separable cases, where true marginals are precisely from the given SAR and SEM specifications, we do indeed see a somewhat better performance for their BMA versions against GPR-BMA, especially for the SAR model where local marginals are directly influenced by $(I_n - \hat{\rho}W)^{-1}$ [as in (64) above]. However, for all nonseparable cases, GPR-BMA is seen to yield a dramatic improvement over both SAR-BMA and SEM-BMA (in the LME-RMSE columns). As mentioned above, this is largely due to the frequent sign changes in LME for the nonseparable case, which cannot be captured by either of these models. It is here that the real strength of GPR-BMA is most evident.

Finally, we note that for this particular nonseparable model in which the true AME values are close to zero, it is difficult to gauge the effectiveness of AME in summarizing the directional effects of variables deemed relevant by GPR-BMA. So to check such effects, we have added two columns to the nonseparable case summarizing the percentages of correct LME signs at the individual observation level. Here, we see that for the relevant variables, $(x_1, x_2, x_3)$, these estimated signs are extremely reliable for GPR-BMA, thus suggesting that AME signs will continue to provide a good summary measure of directional influences in those cases where true AME values are substantially different from zero.

In summary, the single most important result of these simulated comparisons is to underscore the robust performance of GPR-BMA in all cases. Even though there is no attempt to capture either the conditional mean specifications or spatial autocorrelation structures of the models simulated, GPR-BMA identifies the statistically relevant set of variables with striking regularity. Moreover, its performance is strongest in precisely those cases where specification

errors tend to degrade the performance of more parametric models such as SAR-BMA and SEM-BMA. While these results may at first glance appear "too good to be true," they serve to underscore the main difference between global parametric and local nonparametric approaches. By focusing primarily on local information around each location, the latter approach is able to discern changing relationships with a remarkable degree of reliability. Finally, it should also be added that BMA seems to work especially well in this setting. In particular, it effectively dampens variations in these local relationships over the many alternative candidate models in $\mathbb{M}$.[12]

## Empirical application: Predictors of economic growth

In this final section, we apply GPR-BMA to a real data set to highlight how well this technique performs in practice. Here, we use a standard BMA benchmark data set focusing on economic growth (Sala-i-Martin 1997; Fernandez, Ley, and Steel 2001a,b), which includes 42 candidate explanatory variables for each of 72 countries. To capture possible spatial effects, we include the spatial information (absolute latitude[13] and longitude) of each country. As one such example, it has been claimed by Sachs (2001) and others that technology diffuses more readily across the same latitude than the same longitude. Such assertions can be tested within the present framework (as shown below).

Since OLS-BMA is most often used in conjunction with this data set, we provide OLS-BMA estimates alongside results produced by the spatial techniques of SAR-BMA and SEM-BMA.[14] Using OLS-BMA as a benchmark for comparison, GPR-BMA was calibrated to have a prior expected model size equal to the estimated average model size of OLS-BMA. In particular, since expected model size is given by the sum of inclusion probabilities for all variables $[E(q) = E(\Sigma_{j=1}^{k}\delta_j) = \Sigma_{j=1}^{k}E(\delta_j) = \Sigma_{j=1}^{k}p(\delta_j = 1)]$, this realized sum for OLS-BMA ($\approx$ 10.4) is taken as the mean of the prior distribution for $q$ in expression (19) and is used to solve numerically for $\lambda$, yielding a value of $\lambda \approx 0.089$ (as shown in Fig. 1 above).[15] The resulting VIP for all BMA techniques are shown in the first four columns of Table 2 below [where the OLS-BMA estimates are calculated using MATLAB code from Koop, Poirier and Tobias (2007)]. The first two rows include spatial variables, and the remaining rows are ordered by inclusion probabilities under GPR-BMA.

Observe first that there is strong agreement between the sets of inclusion probabilities, with an overall correlation of greater than 80% between GPR-BMA and each of the remaining methods. In particular, these methods are in general agreement as to the most important variables (with inclusion probabilities above 0.90), with the single exception of *Non-Equipment Investment*. Here, the inclusion probability under GPR-BMA (0.948) is more than twice that of any other model. Further investigation suggests that there are collinearities between *Equipment Investment* and *Non-Equipment Investment* (depending on which other explanatory variables are present). Moreover, since the linear specification used in the other three models is well known to be more sensitive to such collinearities than GPR, this could well be the main source of the difference.

Turning next to the spatial variables, absolute latitude, and longitude, it is clear that neither is a relevant predictor of economic growth in the present data set. These values for absolute latitude provide little support for the Sachs (2001) hypothesis. Similar insignificance was obtained when absolute latitude was replaced by latitude (results not shown). For this latter specification, note also from the form of the squared exponential kernel in (2) that the presence of both

**Table 2.** Variable Inclusion Probabilities and Average Marginal Effects

| | VIP | | | | AME | | | |
|---|---|---|---|---|---|---|---|---|
| Variable Name | GPR-BMA | OLS-BMA | SAR-BMA | SEM-BMA | GPR-BMA | OLS-BMA | SAR-BMA | SEM-BMA |
| Absolute (latitude) | 0.210 | 0.038 | 0.011 | 0.010 | 0.000 | 0.000 | 0.000 | 0.000 |
| Longitude | 0.230 | 0.040 | 0.009 | 0.010 | 0.000 | 0.000 | 0.000 | 0.000 |
| ln(GDP) in 1960 | 0.999 | 0.998 | 0.996 | 0.999 | −0.012 | −0.016 | −0.016 | −0.016 |
| Equipment Investment | 0.962 | 0.923 | 0.984 | 0.964 | 0.153 | 0.161 | 0.193 | 0.188 |
| Life Expectancy | 0.959 | 0.932 | 0.935 | 0.954 | 0.001 | 0.001 | 0.001 | 0.001 |
| Nonequipment Investment | 0.948 | 0.423 | 0.242 | 0.262 | 0.046 | 0.024 | 0.014 | 0.015 |
| Fraction Confucian | 0.877 | 0.987 | 0.990 | 0.988 | 0.049 | 0.056 | 0.056 | 0.056 |
| Sub-Saharan Africa | 0.726 | 0.735 | 0.721 | 0.619 | −0.008 | −0.011 | −0.011 | −0.009 |
| Number Of Years Open | 0.674 | 0.517 | 0.559 | 0.641 | 0.007 | 0.007 | 0.009 | 0.010 |
| Age | 0.664 | 0.079 | 0.033 | 0.025 | 0.000 | 0.000 | 0.000 | 0.000 |
| Degree of Capitalism | 0.609 | 0.453 | 0.310 | 0.249 | 0.001 | 0.001 | 0.001 | 0.001 |
| Fraction Hindu | 0.603 | 0.118 | 0.024 | 0.014 | −0.026 | −0.003 | 0.000 | 0.000 |
| Fraction Muslim | 0.580 | 0.636 | 0.597 | 0.681 | 0.005 | 0.009 | 0.008 | 0.009 |
| Rule of law | 0.551 | 0.490 | 0.404 | 0.355 | 0.005 | 0.007 | 0.006 | 0.005 |
| Latin America | 0.496 | 0.209 | 0.127 | 0.104 | −0.004 | −0.002 | −0.001 | −0.001 |
| Fraction Buddhist | 0.445 | 0.196 | 0.068 | 0.083 | 0.007 | 0.003 | 0.001 | 0.001 |
| Fraction Protestants | 0.420 | 0.452 | 0.349 | 0.257 | −0.005 | −0.006 | −0.005 | −0.003 |
| Political Rights | 0.420 | 0.089 | 0.034 | 0.035 | 0.000 | 0.000 | 0.000 | 0.000 |
| Primary Exports | 0.417 | 0.097 | 0.026 | 0.052 | −0.004 | −0.001 | 0.000 | −0.001 |
| % of Pop. Speaking English | 0.401 | 0.070 | 0.016 | 0.027 | −0.003 | 0.000 | 0.000 | 0.000 |
| Ratio Workers to Population | 0.340 | 0.045 | 0.021 | 0.019 | −0.001 | 0.000 | 0.000 | 0.000 |
| Size of Labor Force | 0.337 | 0.077 | 0.013 | 0.019 | 0.000 | 0.000 | 0.000 | 0.000 |
| Fraction of GDP in Mining | 0.332 | 0.451 | 0.267 | 0.291 | 0.004 | 0.019 | 0.011 | 0.013 |

**Table 2** *Continued*

| | VIP | | | | AME | | | |
|---|---|---|---|---|---|---|---|---|
| Variable Name | GPR-BMA | OLS-BMA | SAR-BMA | SEM-BMA | GPR-BMA | OLS-BMA | SAR-BMA | SEM-BMA |
| Black Market Premium | 0.312 | 0.172 | 0.046 | 0.055 | −0.001 | −0.001 | 0.000 | 0.000 |
| Fraction Catholic | 0.307 | 0.128 | 0.074 | 0.020 | −0.001 | 0.000 | 0.000 | 0.000 |
| Civil Liberties | 0.303 | 0.120 | 0.067 | 0.073 | 0.000 | 0.000 | 0.000 | 0.000 |
| Higher Education Enrollment | 0.295 | 0.042 | 0.011 | 0.012 | −0.015 | −0.001 | 0.000 | 0.000 |
| Primary School Enrollment | 0.280 | 0.205 | 0.165 | 0.167 | 0.000 | 0.004 | 0.004 | 0.004 |
| Ethnolinguistic Fractionalization | 0.268 | 0.055 | 0.012 | 0.011 | 0.002 | 0.000 | 0.000 | 0.000 |
| War | 0.230 | 0.073 | 0.032 | 0.030 | −0.001 | 0.000 | 0.000 | 0.000 |
| Spanish Colony | 0.206 | 0.050 | 0.013 | 0.013 | 0.001 | 0.000 | 0.000 | 0.000 |
| Population Growth | 0.193 | 0.038 | 0.011 | 0.010 | −0.008 | 0.005 | 0.002 | 0.002 |
| % of Pop. Speaking Foreign Language | 0.173 | 0.066 | 0.021 | 0.033 | 0.000 | 0.000 | 0.000 | 0.000 |
| French Colony | 0.166 | 0.049 | 0.019 | 0.018 | 0.000 | 0.000 | 0.000 | 0.000 |
| Exchange Rate Distortions | 0.159 | 0.076 | 0.035 | 0.049 | 0.000 | 0.000 | 0.000 | 0.000 |
| Fraction Jewish | 0.146 | 0.036 | 0.011 | 0.009 | 0.001 | 0.000 | 0.000 | 0.000 |
| British Colony | 0.140 | 0.037 | 0.009 | 0.008 | 0.000 | 0.000 | 0.000 | 0.000 |
| St. Dev. of Black Market Premium | 0.137 | 0.050 | 0.018 | 0.024 | 0.000 | 0.000 | 0.000 | 0.000 |
| Outward Orientation | 0.127 | 0.038 | 0.008 | 0.008 | 0.000 | 0.000 | 0.000 | 0.000 |
| Area | 0.121 | 0.028 | 0.009 | 0.008 | 0.000 | 0.000 | 0.000 | 0.000 |
| Revolutions and Coups | 0.111 | 0.029 | 0.011 | 0.009 | 0.000 | 0.000 | 0.000 | 0.000 |
| Public Education Share | 0.101 | 0.028 | 0.007 | 0.008 | 0.006 | 0.001 | 0.000 | 0.000 |

latitude and longitude should, in principle, capture any effects of squared Euclidean (decimal-degree) distances on covariance. So for GPR-BMA in particular, these low inclusion probabilities suggest that there is little in the way of spatial dependency among these national economic growth rates (after controlling for the other explanatory variables).

**Average marginal effects**

To compare the AMEs of variables in GPR-BMA with those of SAR-BMA, SEM-BMA, and OLS-BMA, we calculate these values using the same procedures outlined for the simulation studies above (where the treatment of OLS-BMA in terms of beta estimates is identical to that for SEM-BMA). The estimated AME for each variable is shown in the last four columns of Table 2, where it is again seen that the results for all models are quite similar. This similarity is even stronger when one considers the influence of VIP on marginal effects [as seen for GPR-BMA in expression (53) above]. In particular, differences in AMEs between these methods are often the result of corresponding differences between their associated VIP.

As one example, recall that for *Non-Equipment Investment* the inclusion probability under GPR-BMA is roughly twice that under OLS-BMA. So, given that the AME of this variable in GPR-BMA is also roughly twice that in OLS-BMA, one can conclude that the AME restricted to those models where *Non-Equipment Investment* is present are actually quite similar for these two methods. The same argument holds for comparisons with SAR-BMA and SEM-BMA as well. Note finally that the strong agreement in *signs* of AMEs among all models suggests that for GPR-BMA in particular, these estimates should provide reliable indicators of the average *direction* of marginal effects for all relevant variables.

**Local marginal effects**

While AMEs under GPR-BMA are similar to those under the other models, it is possible to probe deeper with GPR-BMA. For unlike many regression-based models, where the marginal effects of variables are hypothesized to be constant across space, one can "drill down" with GPR-BMA and examine marginal effects at different data locations, such as countries in the present case. Moreover, such local results can, in principle, reveal structural relations between marginal effects and other variables that are not readily accessible by other models. A partial exception here is SAR-BMA, where local marginal effects are to some degree meaningful.[16] While in many empirical contexts both the signs and ratios of such effects among variables are approximately constant across space, it is nonetheless of interest to compare them with those of GPR-BMA, as is done in Table 3 below. In particular, we now consider differences between the marginal effects of *Equipment Investment* and *Non-Equipment Investment* across countries, as displayed in Table 3 (where the marginal effects of *Equipment Investment* are shown in descending order).

Turning first to *Equipment Investment*, we see from the GPR-BMA results that the highest marginal effects on economic growth are exhibited by less developed countries (such as Cameroon) and the lowest marginal effects by more developed countries (such as the United States). In contrast, these LMEs for SAR-BMA are virtually constant across countries. Closer inspection shows that this is due to the weak level of spatial dependence estimated by SAR-BMA (with a model-averaged $\hat{\rho}$ of only 0.04). So while LMEs are indeed present, it is evident that they not attributable to spatial-feedback relations, as hypothesized by SAR-BMA. Rather, they appear to be related to the overall level of economic development in each country, as revealed by the GPR-BMA results.

**Table 3.** Marginal Effect of Equipment and Nonequipment Investment by Country

| | GPR-BMA | | SAR-BMA | | | GPR-BMA | | SAR-BMA | |
|---|---|---|---|---|---|---|---|---|---|
| | Eq. Inv. | NE Inv. | Eq. Inv. | NE Inv. | | Eq. Inv. | NE Inv. | Eq. Inv. | NE Inv. |
| Malawi | 0.222 | 0.094 | 0.193 | 0.014 | Brazil | 0.154 | 0.054 | 0.193 | 0.014 |
| Cameroon | 0.212 | 0.091 | 0.193 | 0.014 | Algeria | 0.153 | 0.012 | 0.193 | 0.014 |
| Tanzania | 0.211 | 0.081 | 0.193 | 0.014 | Panama | 0.153 | 0.040 | 0.193 | 0.014 |
| Kenya | 0.210 | 0.059 | 0.193 | 0.014 | Chile | 0.151 | 0.057 | 0.193 | 0.014 |
| Nigeria | 0.210 | 0.079 | 0.193 | 0.014 | Mexico | 0.148 | 0.047 | 0.193 | 0.014 |
| Ethiopia | 0.209 | 0.115 | 0.193 | 0.014 | India | 0.147 | 0.040 | 0.193 | 0.014 |
| Madagascar | 0.207 | 0.107 | 0.193 | 0.014 | Costa Rica | 0.146 | 0.042 | 0.193 | 0.014 |
| Uganda | 0.206 | 0.105 | 0.193 | 0.014 | Argentina | 0.144 | 0.068 | 0.193 | 0.014 |
| Zaire | 0.205 | 0.105 | 0.193 | 0.014 | Taiwan | 0.142 | 0.044 | 0.193 | 0.014 |
| Zimbabwe | 0.200 | 0.067 | 0.193 | 0.014 | Portugal | 0.142 | 0.038 | 0.193 | 0.014 |
| Ghana | 0.200 | 0.090 | 0.193 | 0.014 | Uruguay | 0.140 | 0.055 | 0.193 | 0.014 |
| Congo | 0.199 | 0.052 | 0.193 | 0.014 | Venezuela | 0.137 | 0.041 | 0.193 | 0.014 |
| Senegal | 0.195 | 0.085 | 0.193 | 0.014 | Spain | 0.135 | 0.032 | 0.193 | 0.014 |
| Zambia | 0.186 | 0.007 | 0.193 | 0.014 | Cyprus | 0.130 | 0.002 | 0.193 | 0.014 |
| Philippines | 0.186 | 0.084 | 0.193 | 0.014 | Greece | 0.127 | −0.001 | 0.193 | 0.014 |
| Pakistan | 0.185 | 0.070 | 0.193 | 0.014 | United Kingdom | 0.126 | 0.038 | 0.193 | 0.014 |
| Haiti | 0.182 | 0.102 | 0.193 | 0.014 | South Korea | 0.126 | 0.039 | 0.193 | 0.014 |
| Thailand | 0.175 | 0.061 | 0.193 | 0.014 | Ireland | 0.125 | 0.008 | 0.193 | 0.014 |
| Morocco | 0.174 | 0.075 | 0.193 | 0.014 | Hong Kong | 0.122 | 0.044 | 0.193 | 0.014 |
| Bolivia | 0.172 | 0.066 | 0.193 | 0.014 | Italy | 0.119 | 0.009 | 0.193 | 0.014 |
| Honduras | 0.171 | 0.056 | 0.193 | 0.014 | Denmark | 0.116 | 0.018 | 0.193 | 0.014 |
| Tunisia | 0.170 | 0.063 | 0.193 | 0.014 | Belgium | 0.113 | 0.005 | 0.193 | 0.014 |
| Sri Lanka | 0.169 | 0.064 | 0.193 | 0.014 | Australia | 0.112 | 0.007 | 0.193 | 0.014 |
| El Salvador | 0.167 | 0.082 | 0.193 | 0.014 | Sweden | 0.112 | −0.002 | 0.193 | 0.014 |
| Turkey | 0.167 | 0.052 | 0.193 | 0.014 | Austria | 0.110 | 0.040 | 0.193 | 0.014 |
| Paraguay | 0.165 | 0.082 | 0.193 | 0.014 | Canada | 0.109 | 0.022 | 0.193 | 0.014 |
| Guatemala | 0.165 | 0.054 | 0.193 | 0.014 | Israel | 0.108 | 0.017 | 0.193 | 0.014 |
| Dominican Republic | 0.165 | 0.064 | 0.193 | 0.014 | Germany | 0.107 | −0.004 | 0.193 | 0.014 |
| Peru | 0.164 | 0.073 | 0.193 | 0.014 | Netherlands | 0.107 | 0.004 | 0.193 | 0.014 |
| Nicaragua | 0.164 | 0.048 | 0.193 | 0.014 | United States | 0.106 | 0.022 | 0.193 | 0.014 |
| Botswana | 0.164 | 0.046 | 0.193 | 0.014 | France | 0.105 | 0.004 | 0.193 | 0.014 |
| Jordan | 0.162 | 0.044 | 0.193 | 0.014 | Switzerland | 0.105 | −0.002 | 0.193 | 0.014 |
| Malaysia | 0.160 | 0.023 | 0.193 | 0.014 | Norway | 0.102 | −0.004 | 0.193 | 0.014 |
| Colombia | 0.159 | 0.059 | 0.193 | 0.014 | Finland | 0.097 | −0.020 | 0.193 | 0.014 |
| Jamaica | 0.155 | 0.045 | 0.193 | 0.014 | Japan | 0.083 | −0.001 | 0.193 | 0.014 |
| Ecuador | 0.155 | 0.027 | 0.193 | 0.014 | Singapore | 0.069 | 0.023 | 0.193 | 0.014 |

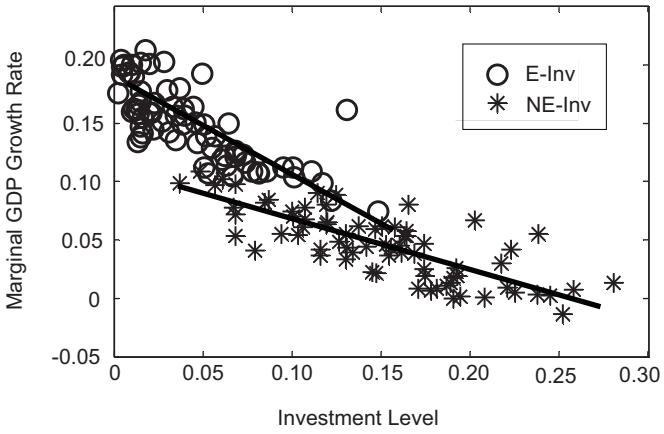


**Figure 2.** Investment comparisons.

Using these marginal results one can discern an interesting relation between *Equipment Investment* and *Non-Equipment Investment* by plotting their marginal effects against corresponding investment levels for each country, as shown in Fig. 2 (where circles and stars are used to represent marginal effects *Equipment Investment* and *Non-Equipment Investment*, respectively).

Here, the negative slopes of both sets of values suggest that both types of investments exhibit diminishing returns with respect to their marginal effects on economic growth. In addition, this plot also suggests that economic growth is more sensitive to changes in *Equipment Investment* than other types of investment. Both of these observations are easily quantified by regressing marginal effects on investment levels together with a categorical investment-type variable and interaction term. These regression results (not reported) show that both observations above are strongly supported, and in particular, that the response slope for *Equipment Investment* is indeed much steeper than that for other investments (as seen graphically by the regression lines plotted in Fig. 2). In summary, such results serve to illustrate how GPR-BMA can be used to address a wide range of questions not accessible by more standard regression-based approaches.

## Concluding remarks

The objective of this article has been to develop Gaussian Process Regression with Bayesian Model Averaging (GPR-BMA) as an alternative tool for spatial data analysis. This method combines the predictive capabilities of nonparametric methods with many of the more explanatory capabilities of parametric methods. Here, our main effort has been to show by means of selected simulation studies that this method can serve as a powerful exploratory tool when little is known about the underlying structural relations governing spatial processes. Our specific strategy has been to focus on the simplest types of nonseparable relations beyond the range of standard exploratory linear regression specifications, and to show that with only a minimum number of parameters, GPR-BMA is able to identify not only the relevant variables governing such relations, but also the local marginal effects of such variables.

As noted in the "Simulated model comparisons" Section above, it is in principle possible to construct sufficiently elaborate specifications of parametric regressions that will also identify

the particular nonseparable relationships used here, or indeed almost any type of relationship. But it must be stressed that the introduction of such "contingent interaction" parameters requires large sample sizes and tends to suffer from over-fitting problems. Alternatively, one can capture such relationships by directly parameterizing local marginal effects themselves, as in local linear regression methods such as Geographically Weighted Regression. But while such "nonparametric" methods are indeed better able to capture local variations in relationships, they do so by in fact introducing a host of local regression parameters that are highly susceptible to collinearity problems (not to mention the need for exogenously specified bandwidth parameters that are essential for spatially weighted regressions).[17] Moreover, the focus of these models on local effects of variables tends to ignore the possible global relations among them.[18] So, the main result of our simulations is to show that by modeling covariance relations rather than conditional means, the simple version of GPR-BMA developed here is able to identify complex relationships with only *three* model parameters. This is in part explained by the general robustness properties of BMA. But as we have seen in both SAR-BMA and SEM-BMA, such model averaging by itself may not be very effective when unmodeled nonseparabilities are present. So, an important part of the explanation for the success of present GPR-BMA model appears to be the ability of the squared-exponential covariance kernel in GPR to capture both global and local interactions in terms of its scale and bandwidth parameters, $v$ and $\tau$.

This ability to capture both global and locaractions has a wide range of applications in empirical analyses, as in our economic growth example. Here, we saw that GPR-BMA was not only able to capture global determinants of economic growth in a manner similar to SAR-BMA, SEM-BMA, and OLS-BMA, but was also able to delve deeper. In particular, the LMEs of investment estimated by GPR-BMA (across countries) were used to obtain evidence for diminishing returns to investment, and in particular, for stronger diminishing returns with respect to equipment investment.

But in spite of these advantages, it must also be emphasized that the parsimonious parameterization of the present GPR-BMA model is only made possible by the underlying assumptions of *zero means* together with both *stationarity* and *isotropy* of the covariance kernel. While the zero-mean and isotropy assumptions have been mollified to a certain degree by the use of standardized variables, it is nonetheless of interest to consider extensions of the present model that avoid the need for such artificial standardizations. For example, as we have already seen in expression (17) above, extended parameterizations are possible in which individual bandwidths are assigned to each parameter. In addition, it is possible to relax the zero mean assumption by internally estimating a constant mean, $\mu(x) = \mu$, in expression (1) or even by modeling means as parameterized functions of $x$ (as for example in Section 2.7 of Rasmussen and Williams 2006). But a key point to bear in mind here is that the important *conditional means* in expression (8) are much less sensitive to such specifications that the overall Gaussian process itself.

Perhaps the most interesting extensions of the present model are in terms of possible relaxations of the covariance stationarity assumption (which cannot be mollified by any simple standardization procedures). A number of extensions along these lines have been proposed that amount to partitioning space into regions that are approximately stationary, and patching together appropriate covariance kernels for each region. The most recent contribution along these lines appears to be the work of Konomi, Sang, and Mallick (2013), in which regression-tree methods are used for adaptively partitioning space, and in which covariance kernels are

constructed using the "full approximation" method of Sang and Huang (2012). Adaptations of such schemes to the present GPR-BMA framework will be explored in subsequent work.

In addition to these structural assumptions, the single strongest limitation of the present GPR-BMA model is the scaling of its computation time with respect to the number of observations. This is an active area of research where a variety of methods having been proposed over the past few years. Generally speaking, most approaches recommend some type of data reduction technique (see Cornford, Csató, and Opper (2005) for an early example). Solutions range from direct subsampling of the data itself to more sophisticated constructions of "best representative" virtual data set (as compared in detail by Chalupka, Williams, and Murray 2013). Alternative approaches have been proposed that involve lower dimensional approximations to covariance kernels, as in the recent the "random projection" method of Banerjee, Dunson, and Tokdar (2013). But for our purposes, data reduction methods have the advantage of allowing our BMA methods to be preserved intact.

In conclusion, while much work remains to be done in this burgeoning field, our own next steps will be to explore methods for increasing the computational efficiency GPR-BMA in a manner that broadens its range of applications. Our particular focus will be on richer covariance structures that can capture both anisotropic and nonstationary phenomena. For example, by relaxing the present isotropy assumption and using different length scales for latitude and longitude, we can in principle sharpen our test of Sach's (2001) hypothesis discussed in the empirical applications section. Such extensions will be reported in a subsequent paper.

## Notes

1 For an overview of nonparametric inductive approaches to spatial data analysis, see for example, Gahegan (2000).

2 For an overview of alternative "filtering" approaches to spatial regression, see Getis and Griffith (2002).

3 A number of local regression approaches are also capable of both prediction and variable identification (Brunsdon, Fotheringham, and Charlton 1996; McMillen 1996). However, due to space limitations, a systematic comparison with these nonparametric methods is deferred to a subsequent paper.

4 As discussed in BMA approach Section below, our present formulation differs slightly from Chen and Wang (2010) in terms of model sampling.

5 For alternative approach using Monte Carlo methods in the context of spatial kriging with location uncertainty, see Gabrosek and Cressie (2002).

6 As pointed out by Chen and Wang, this independence assumption greatly simplifies the MCMC analysis to follow. In particular, if covariance functions such as (17) are used, then the parameter vector $\theta$essentially changes dimension with each model. This requires more complex reversible-jump methods (Green 1995) that tend to be computationally intensive. So as stated previously, our objective is to show that even without such refinements, the present GPR-BMA procedure performs remarkably well.

7 Note that in principle it is also possible to analyze marginal effects on $E(y_l|x_l, \tilde{y}, \tilde{X}, \delta_i, \theta_i)$ with respect to changes in explanatory variables, $\tilde{x}_{sj}$, at data locations, $s$. In this context, it can be seen that the inverse $K_{\theta_i}[\tilde{X}(\delta_i)]^{-1}$, in (54) plays a role similar to the "indirect effects" induced by the inverse $(I_n - \rho W)^{-1}$ in (60) below for the SAR model (as brought to our attention by a referee, and developed in detail by LeSage and Pace 2009, Section 2.7.1). However, we shall not pursue such indirect marginal effects in this article.

8 Here, it is important to note that average parameter values are not directly available in the LeSage and Parent formulation of SAR-BMA and SEM-BMA (where such information is integrated out to construct posterior model probabilities). Thus, it was necessary to augment their approach to obtain such information. The strategy used here was first to calculate parameter estimates by running SAR

(respectively, SEM) for each model found by SAR-BMA (respectively, SEM-BMA), and then to average these estimates across models weighted by their posterior model probabilities.

9 Note that expressions such as $x_1 \cdot x_2$ for vectors are implicitly component-wise (Hadamard) products.

10 However, as with all Bayesian methods, GPR-BMA does require prior specifications of certain parameters: in this case, the hyperparameters in (11) and the tuning parameter, $\lambda = 0.01$, in (19). In addition, certain computational conventions are used: here including a "jitter" of size 0.02 added to $\sigma$ in (5) for numerical stability of inverses, as well as MCMC conventions such as our burn-in threshold of 500 iterations.

11 Low VIPs for the full set of explanatory variables imply that the *null model* (intercept only) is very frequent, and has all zero marginals by definition.

12 For brevity's sake, we do not pursue questions of out-of-sample performance in this article. Elsewhere, we have addressed the issue of out-of-sample prediction where we have found that GPR-BMA generates very accurate out-of-sample forecasts even when faced with highly nonlinear (and unknown) relationships. A full discussion of the out-of-sample performance of GPR-BMA is available from the authors.

13 Fernandez, Ley, and Steel (2001a,b) and Sala-i-Martin (1997) use absolute latitude to distinguish between tropic and temperate zones. We follow their approach to maintain consistency with previous research.

14 The spatial weights matrix, $W$, used for both SAR-BMA and SEM-BMA, was here taken to be a standard contiguity matrix between countries (normalized to have unit maximum eigenvalue).

15 We initialize the first model vector based on the length scales produced by ARD. In particular, we include the first 10 variables based on the shortest (and hence relatively most important) length scales. We further parameterize the model by selecting a "jitter" of $1 \times 10^{-6}$ (for numerical stability) and a burn-in of 1,000.

16 As mentioned in the SAR-BMA and SEM-BMA models Section, local differences of marginal effects among locations, $i$, are here embodied in a spatial multiplier, $b_{\rho}^{ii}$, reflecting all feedback effects at $i$ resulting from the spatial relations implicit in both $W$ and $\rho$ (as detailed in LeSage and Pace 2009, Section 2.7.1).

17 Due to space limitations, a systematic comparison between GPR-BMA and Geographically Weighted Regression is deferred to a subsequent paper.

18 While there are indeed "mixed" versions of such models that incorporate both global (parametric) and local (nonparametric) specifications (as detailed, e.g., in Wei and Qi 2012; Mei, Wang, and Zhang 2006; and in Chapter 3 of Fotheringham, Brunsdon, and Charlton 2002), such models involve a prior partitioning of these variable types, so that no variable is treated both globally and locally.

## References

Banerjee, A., D. B. Dunson, and S. T. Tokdar. (2013). "Efficient Gaussian Process Regression for Large Data Sets." *Biometrika* 100, 75–89.

Bivand, R. S., V. Gómez-Rubio, and H. Rue. (2014). "Approximate Bayesian Inference for Spatial Econometrics Models." *Spatial Statistics* 9, 146–65.

Brunsdon, C., A. S. Fotheringham, and M. C. Charlton. (1996). "Geographically Weighted Regression." *Geographical Analysis* 28, 281–98.

Chalupka, K., C. Williams, and I. Murray. (2013). "A Framework for Evaluating Approximation Methods for Gaussian Process Regression." *Journal of Machine Learning Research* 14, 333–50.

Chen, T., and B. Wang. (2010). "Bayesian Variable Selection for Gaussian Process Regression: Application to Chemometric Calibration of Spectrometers." *Neurocomputing* 73, 2718–26.

Cornford, D., L. Csató, and M. Opper. (2005). "Sequential, Bayesian Geostatistics: A Principled Method for Large Data Sets." *Geographical Analysis* 37(2), 183–99.

Dearmon, J. and T. E. Smith. (2014). "Gaussian Process Regression and Bayesian Model Averaging: An alternative approach to modeling spatial phenomena", http://www.seas.upenn.edu/~tesmith/DEARMON_PAPER.pdf.

Denison, D. G. T., B. K. Mallick, and A. F. M. Smith. (1998). "Bayesian MARS." *Statistics and Computing* 8, 337–46.

Duane, S., A. Kennedy, B. Pendleton, and D. Roweth. (1987). "Hybrid Monte Carlo." *Physics Letters B* 195, 216–22.

Fernandez, C., E. Ley, and M. Steel. (2001a). "Model Uncertainty in Cross-Country Growth Regressions." *Journal of Applied Econometrics* 16, 563–76.

Fernandez, C., E. Ley, and M. Steel. (2001b). "Benchmark Priors for Bayesian Model Averaging." *Journal of Econometrics* 100, 381–427.

Fotheringham, A. S., C. Brunsdon, and M. Charlton. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. New York: Wiley.

Gabrosek, J., and N. Cressie. (2002). "The Effect on Attribute Prediction of Location Uncertainty in Spatial Data." *Geographical Analysis* 34(3), 262–85

Gahegan, M. (2000). "On the Application of Inductive Machine Learning Tools to Geographical Analysis." *Geographical Analysis* 32(2), 113–39.

George, E. I., and R. E. McCulloch. (1993). "Variable Selection via Gibbs Sampling." *Journal of the American Statistical Association* 88, 881–9.

Getis, A., and D. A. Griffith. (2002). "Comparative Spatial Filtering in Regression Analysis." *Geographical Analysis* 34(2), 130–40.

Green, P. J. (1995). "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination." *Biometrika* 82, 711–32.

Konomi, B. A., H. Sang, and B. K. Mallick. (2013). "Adaptive Bayesian Nonstationary Modeling for Large Spatial Datasets using Covariance Approximations." *Journal of Computational and Graphical Statistics* 23, 802–29. DOI 10.1080/10618600.2013.812872

Koop, G., D. J. Poirier, and J. L. Tobias. (2007). *Bayesian Econometric Methods*, Vol. 7. Cambridge, England: Cambridge University Press.

LeSage, J., and M. Fischer. (2008). "Spatial Growth Regressions: Model Specification, Estimation and Interpretation." *Spatial Economic Analysis* 3, 275–304.

LeSage, J., and R. K. Pace. (2009). *Introduction to Spatial Econometrics*. Boca Raton, FL: Chapman-Hall.

LeSage, J., and O. Parent. (2007). "Bayesian Model Averaging for Spatial Econometric Models." *Geographical Analysis* 39, 241–67.

MacKay, D. J. (1995). "Probable Networks and Plausible Predictions - A Review of Practical Bayesian Methods for Supervised Neural Networks." *Network: Computation in Neural Systems* 6(3), 469–505.

MacKay, D. J. C. (1998). "Introduction to Gaussian Processes." In *Neural Networks and Machine Learning*, 133–65, edited by C. M. Bishop. Berlin: Springer.

McMillen, D. (1996). "One Hundred Fifty Years of Land Values in Chicago: A Nonparametric Approach." *Journal of Urban Economics* 40, 100–24.

Mei, C.-L., N. Wang, and W.-X. Zhang. (2006). "Testing the Importance of the Explanatory Variables in a Mixed Geographically Weighted Regression." *Environment and Planning A* 38, 587–98.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks, Lecture Notes in Statistics 118*. New York: Springer.

Neal, R. M. (2010). "MCMC Using Hamiltonian Dynamics." In *Chapter 5 in Handbook of Markov Chain Monte Carlo*, edited by S. Brooks, A. Gelman, G. Jones, and X. L. Meng. London: Chapman and Hall.

Raftery, A., D. Madigan, and J. Hoeting. (1997). "Bayesian Model Averaging for Linear Regression Models." *Journal of the American Statistical Association* 92, 179–91.

Rasmussen, C., and C. Williams. (2006). *Gaussian Process for Machine Learning*. Cambridge MA: MIT Press.

Sachs, J. (2001). "Tropical Underdevelopment." *Working Paper 8119*. NBER Working Paper Series, National Bureau of Economic Research, Cambridge, MA.

Sala-i-Martin, X. X. (1997). "I Just Ran Two Million Regressions." *The American Economic Review* 87, 178–83.

Sang, H., and J. Z. Huang. (2012). "A Full Scale Approximation of Covariance Functions for Large Data Sets." *Journal of the Royal Statistical Society* 74, 111–32.

Shi, J. Q., and T. Choi. (2011). *Gaussian Process Regression Analysis for Functional Data*. Boca Raton, FL: CRC Press.

Wei, C.-H., and F. Qi. (2012). "On the Estimation and Testing of Mixed Geographically Weighted Regression *Model*." *Economic Modelling* 29, 2615–20.

Williams, C. K. I., and C. E. Rasmussen. (1996). "Gaussian Processes for Regression." In *Advances in Neural Information Processing Systems 8*, 514–20, edited by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo. Boston: MIT Press.

## Supporting Information

Additional Supporting Information may be found in the online version of this article.