Gaussian Process Regression and Bayesian Model Averaging: An alternative approach to modeling spatial phenomena

Jacob Dearmon
Meinders School of Business
Oklahoma City University

Tony E. Smith
Department of Electrical and Systems Engineering
University of Pennsylvania

July 20, 2014

Abstract

Gaussian Process Regression (GPR) is an estimation 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 (SAR) models or spatial error models (SEM), place rather strong prior restrictions on the functional form of relationships, but allow direct inference with respect to explanatory variables. In this paper, we attempt to combine the best of both techniques by augmenting GPR with a Bayesian Model Averaging (BMA) component which allows for the identification of statistically relevant explanatory variables while retaining the predictive performance of GPR.

Other approaches along these lines include the well-known BMA extensions of both SAR and SEM, as well as the class of locally weighted regression methods exemplified by Geographically Weighted Regression (GWR). To demonstrate the relative effectiveness of GPR-BMA, we construct several simulated comparisons designed to capture the types of non-separable relationships that are most difficult to identify by standard regression methods. In particular, a simulated spatial housing-price example is constructed that is sufficiently rich to demonstrate the behavioral relevance of such non-separabilities, as well as to allow a wide range of comparisons among these methods. In addition, we also apply GPR-BMA to a benchmark BMA dataset on economic growth to illustrate certain additional insights made possible by this approach. Our main results show that GPR-BMA not only exhibits better predictive power than these alternative models, but also more accurately identifies the true variables associated with the underlying data generating process. 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. Moreover, while such frequencies offer no direct information about the signs of local marginal effects, it is shown that partial derivatives based on mean GPR predictions do provide such information, and in a manner that exhibits better small-sample properties than GWR.

1. 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 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 paper 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).

But before doing so, we must stress that there have been many attempts achieve such a unification. On the parametric side, the work most closely related to our present approach has been the efforts of LeSage et al. (2007, 2008) to achieve more robust spatial regression models by combining them with Bayesian Model Averaging.[2] On the nonparametric side, a number of methods for variable identification have been proposed for the important class of Locally Weighted Regression (LWR) models. In particular, McMillen et al. (1996, 2010, 2012) have extended the general testing procedures of Cleveland and Devlin (1988) to explicit spatial contexts, and following Robinson (1988), have developed new semiparametric testing procedures for the identification of key explanatory variables. But our own work is more closely related to the many efforts to introduce variable identification into Gaussian Process Regression. Perhaps the most widely known method is Automatic Relevance Determination (ARD), first introduced by Neal (1996) and MacKay (1998). But 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 Bayesian Model Averaging for both prediction and variable identification in GPR models. The key feature of this approach is to allow uncertainties with respect to both relevant explanatory variables and spatial predictions to be treated explicitly.

Hence the main contributions of the present paper are to develop this GPR-BMA method in detail, and to present a series of systematic comparisons of this method with the alternative approaches mentioned above, both in term of simulated and empirical data sets. The simulated data sets are designed to capture the types of nonseparable relationships that are most difficult for standard spatial regression models to detect. Here it is shown that even for the more robust BMA versions of spatial regression models, such relationships continue to be elusive and often yield misleading results. In this regard, locally weighted regressions appear to fare much better [see for example the

[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).

discussion in Fotheringham and Brundson's (1999)]. Our present analysis focuses on Geographically Weighted Regression (GWR) which is currently the most widely used version of LWR in spatial applications [as for example in the ArcGIS implementation based on Fotheringham, Brunsdon, and Charlton (2002)]. But while such models are sufficiently flexible to detect locally varying relationships among variables, they are far less reliable than GPR-BMA in terms of root mean squared error. This is most dramatic in terms of their predictive capabilities. In addition, it should also be emphasized that, unlike GWR models which depend on a auxiliary spatial-kernel and window-size parameters (as well as a multitude of locally estimated beta parameters), we have chosen the simplest possible GPR model with a standard squared-exponential covariance kernel involving only three distinct parameters. Our objective in doing so is to show that even without extensive parameterization, GPR-BMA models can achieve a remarkable degree of model flexibility.

To develop these results, we begin in Section 2 below with a detailed development of the GPR-BMA model. This is followed in Section 3 and 4 with the simulated comparisons between GPR-BMA and the alternative approaches outlined above. Finally, in Section 5, we apply GPR-BMA to an empirical data set involving economic growth rates among countries.

2. Gaussian Process Regression with Bayesian Model Averaging

In this section we develop our proposed methodological procedure for spatial data analysis. This begins in Section 2.1 with a general development of Gaussian processes in a Bayesian setting that focuses on Gaussian Process Regression (GPR) – which amounts to posterior prediction within this framework. The Bayesian Model Averaging (BMA) approach to GPR is then developed in Section 2.2.

2.1. 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, $\mathrm{X} \subseteq \mathbb{R}^k$, then this relationship can be formalized as a stochastic process, $\{y(x) : x \in \mathrm{X}\}$, on 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, $\{y(x) : x \in \mathrm{X}\}$, 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*, $\mathrm{cov}(x, x')$, $x, x' \in \mathrm{X}$, usually written more compactly as

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

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, $\mu(x) = 0$, $x \in \mathrm{X}$, and that the covariance function has some specific parametric form, $\mathrm{cov}(x, x') = c_\omega(x, x')$, designated as the *kernel function* for the process with (hyper)parameter vector, ω. While there are many choices for kernels, one of the simplest and most popular is the *squared exponential kernel*,

$$(2) \qquad \mathrm{c}_\omega(x, x') = v \exp\left(-\tfrac{1}{2\tau^2} \| x - x' \|^2\right) = v \exp\left[-\tfrac{1}{2\tau^2} \sum_{j=1}^{k} (x_j - x'_j)^2\right]$$

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 X, the prior distribution of the associated random vector $y = y(X) = [y(x_i) : i = 1,..,N]$ is assumed to be *multinormal*:

$$(3) \qquad y(X) \sim N\left[0_N, c_\omega(X, X)\right]$$

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 *Gaussian Process Regression* (GPR) is to derive posterior predictions about unobserved y values given observed values (data) at some subset of locations in 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) [RW], we assume that for any realized value, $y(x)$, of the process at $x \in \mathrm{X}$, the associated *observed value*, $\tilde{y}(x)$, is a random variable of the form:

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

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 $\tilde{X} = (\tilde{x}_i : i = 1,..,n) \subset \mathrm{X}$, we seek to predict the unobserved value, $y(x)$, at $x \in \mathrm{X}$. To develop this prediction problem statistically, observe first from (3) and (4) that $\tilde{y}$ is multinormally distributed as

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

Hence, by a second application of (3), it follows that the prior distribution of $(y,\tilde{y})$ must be jointly multinormally distributed as (see for example expression (2.21) in [RW]),

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

Thus, by standard arguments (for example expression (A.6), p.200 in [RW]), one may conclude that the *conditional* distribution $y(x)$ given $(\tilde{y},\tilde{X})$, is of the form

(7) $y \mid x, \tilde{y}, \tilde{X} \sim N\left[E(y \mid x, \tilde{y}, \tilde{X}), \text{var}(y \mid x, \tilde{y}, \tilde{X})\right]$

where by definition,

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

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

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 paper [DS] available from the authors on request).

Up to this point, we have implicitly treated the parameters (ν,τ,σ^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 Markov Chain Monte Carlo (MCMC) methods for Bayesian estimation allow model averaging methods to be used for both variable selection and parameter estimation. For purposes of this paper, we adopt the approach developed in Chen and Wang [CW] (2010).[3]

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

(10) $\theta = (\omega,\sigma^2) = (\nu,\tau,\sigma^2) = (\theta_1,\theta_2,\theta_3)$

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

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

(11) $\ln(\theta_i) \sim N(-3,9)$, $i = 1,2,3$

[As is well known, 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 θ can be written as

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

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

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

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

(14) $p(\tilde{y} \mid \tilde{X}, \theta) = (2\pi)^{-n/2} \det[K_\theta(\tilde{X})]^{-1/2} \exp[-\tfrac{1}{2} \tilde{y}' K_\theta(\tilde{X})^{-1} \tilde{y}]$

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

(15) $p(\theta \mid \tilde{y}, \tilde{X}) p(\tilde{y} \mid \tilde{X}) = p(\theta, \tilde{y} \mid \tilde{X}) = p(\tilde{y} \mid \tilde{X}, \theta) p(\theta \mid \tilde{X}) = p(\tilde{y} \mid \tilde{X}, \theta) p(\theta)$

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

(16) $p(\theta \mid \tilde{y}, \tilde{X}) \propto p(\tilde{y} \mid \tilde{X}, \theta) p(\theta)$

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

2.2 Model and Variable Selection in Gaussian Process Regression

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 in Section 5), 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, $E(y \mid x) = \beta_0 + \Sigma_{j=1}^{k} \beta_j x_j$]. 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 *Automatic Relevance Determination* (ARD) [see for example MacKay (1995, 1998) and Neal (1996) together with the discussions in [RW, 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 τ parameters for each variable,

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

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 τ_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 τ_j by (empirical Bayes) maximum likelihood, and then determine some threshold value, τ_0, for $\hat{\tau}_j$ above which x_j is deemed to be irrelevant for prediction.

2.2.1 Bayesian Model Averaging Approach

In contrast to these variable-selection procedures 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, δ_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 δ 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, θ, to include this model vector, (θ,δ), and proceed to develop an appropriate prior distribution for δ on Δ. 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, δ), we can assume that the priors on θ and δ are *statistically independent*.[4]

To construct a prior distribution for δ, 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 δ can be written as

[4] As pointed out by [CW], this independence assumption greatly simplifies the MCMC analysis to follow. In particular, if covariance functions such as (17) are used, then the parameter vector θ 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.

(18) $p(\delta) = p[\delta, s(\delta)] = p(\delta, q) = p(\delta \mid q)p(q)$

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 [CW] we employ a truncated geometric distribution for q given by

(19) $p(q) = \dfrac{\lambda(1-\lambda)^{q-1}}{1-(1-\lambda)^{k}}$, $q = 1,..,k$

where $\lambda \in (0,1)$. This family of distributions always places more weight on smaller values of q, as is seen in Figure 1 below for selected values of λ with $k = 42$. While [CW] suggest that the (hyper)parameter, λ, 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.

Figure 1

To complete the specification of $p(\delta)$ we assume that $p(\delta \mid q)$ 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

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

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

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

Following [CW], this joint posterior is estimated by Gibbs sampling using the conditional distributions,

(22) $p(\theta \mid \delta, \tilde{y}, \tilde{X}) \propto p(\tilde{y} \mid \tilde{X}, \theta, \delta)p(\theta)$, and

(23) $p(\delta \mid \theta, \tilde{y}, \tilde{X}) \propto p(\tilde{y} \mid \tilde{X}, \theta, \delta)p(\delta)$

We now consider each of these Gibbs steps in turn.

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

(24) $p(\theta_i \mid \theta_{-i}, \delta, \tilde{y}, \tilde{X}) = \dfrac{p(\theta \mid \delta, \tilde{y}, \tilde{X})}{p(\theta_{-i} \mid \delta, \tilde{y}, \tilde{X})} \propto p(\theta \mid \delta, \tilde{y}, \tilde{X}) \propto p(\tilde{y} \mid \theta, \delta, \tilde{X})p(\theta_i)p(\theta_{-i})$

$$\propto p(\tilde{y} \mid \theta, \delta, \tilde{X})p(\theta_i) \quad , \quad i = 1,2,3$$

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 [CW] 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, ρ. 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 δ Posterior. In sampling from the posterior distribution of the model vector, δ, we again follow [CW] 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 [CW], 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 Δ_{q+1} that differ from δ^q by only one component, i.e.,

$$(25) \qquad \Delta_{q+1}(\delta^q) = \left\{ \delta^{q+1} \in \Delta_{q+1} : \sum_{i=1}^{k} | \delta_i^{q+1} - \delta_i^q | = 1 \right\} , \quad q = 1,..,k-1$$

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

$$(26) \qquad \Delta_{q-1}(\delta^q) = \left\{ \delta^{q-1} \in \Delta_{q-1} : \sum_{i=1}^{k} | \delta_i^{q} - \delta_i^{q-1} | = 1 \right\} , \quad q = 2,..,k$$

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

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

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

(28) $$T(\delta \mid \delta^q) = p_r(\delta \mid \delta^q)\, p_a(\delta \mid \delta^q) \ ,$$

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

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

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*, $\pi(\cdot \mid \delta^q)$, can be defined as,

(30) $$\pi(b \mid \delta^q) = \begin{cases} 1 & , q = 1 \\ \frac{1}{2} & , 1 < q < k \\ 0 & , q = k \end{cases} \quad , \text{ and}$$

(31) $$\pi(d \mid \delta^q) = \begin{cases} 0 & , q = 1 \\ \frac{1}{2} & , 1 < q < k \\ 1 & , q = k \end{cases}$$

so that by definition, $\pi(b \mid \delta^q) + \pi(d \mid \delta^q) = 1$ 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 $p_r(\delta \mid b, \delta^q)$ 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 δ^q, it follows that

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

Similarly, if $p_r(\delta \mid d, \delta^q)$ 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 δ^q, it also follows that

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

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

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

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

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

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

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

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

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

Gibbs Sampling. The basic Gibbs sampling procedure outlined above was programmed in Matlab (and is described in more detail in the appendix of [DS]). The M-H algorithm for sampling model vectors, δ, forms the outer loop of this procedure, and the HMC procedure for sampling parameter vectors, θ, 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. For the housing price example below, the average number of post burn-in runs required for convergence across 10 simulations at various samples sizes is 262.[5]

2.2.2 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, δ, given the observed data $(\tilde{y}, \tilde{X})$. In particular, these *model probabilities* are simply the marginal probabilities,

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

[5] Specific Gibb sampling parameter values for the house price example are provided in footnote 14.

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

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

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

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

Hence, assuming approximate independence of the samples, $[(\delta_i,\theta_i) : i = 1,..,N]$, an application of the Law of Large Numbers shows that the *sample average*,

$$(41) \qquad \hat{p}(\delta_\alpha \mid \tilde{y}, \tilde{X}) = \frac{1}{N}\sum_{i=1}^{N} I_\alpha(\delta_i,\theta_i)$$

yields a consistent estimate of each model probability, $p(\delta_\alpha \mid \tilde{y}, \tilde{X})$. But since the number of occurrences of δ_α in the sample sequence, $[(\delta_i,\theta_i) : i = 1,..,N]$, is given by,

$$(42) \qquad N(\delta_\alpha) = \sum_{i=1}^{N} I_\alpha(\delta_i,\theta_i)$$

it follows from (41) that for any model, $\delta \in \Delta$, this estimator is simply the fraction of δ occurrences, i.e.,

$$(43) \qquad \hat{p}(\delta \mid \tilde{y}, \tilde{X}) = \frac{N(\delta)}{N}$$

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

$$(44) \qquad \hat{\delta} = \arg\max_{\delta \in \Delta} \hat{p}(\delta \mid \tilde{y}, \tilde{X}) = \arg\max_{\delta \in \Delta} \frac{N(\delta)}{N}$$

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

$$(45) \qquad p(\delta_j = 1 \mid \tilde{y}, \tilde{X}) = \sum_{\delta \in \Delta_j} p(\delta \mid \tilde{y}, \tilde{X})$$

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

$$(46) \qquad \hat{p}(\delta_j = 1 \mid \tilde{y}, \tilde{X}) = \sum_{\delta \in \Delta_j} \hat{p}(\delta \mid \tilde{y}, \tilde{X}) = \sum_{\delta \in \Delta_j} \frac{N(\delta)}{N}$$

Finally, since

$$(47) \qquad N_j = \sum_{\delta \in \Delta_j} N(\delta)$$

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,

$$(48) \qquad \hat{p}(\delta_j = 1 \mid \tilde{y}, \tilde{X}) = \frac{N_j}{N}$$

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, $(\tilde{y}, \tilde{X})$.

2.2.3 Prediction and Marginal Effects Using Bayesian Model Averaging

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, $x = (x_1,..,x_k) \in \mathrm{X}$, 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 by 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,

$$(49) \quad E[y \mid x(\hat{\delta}), \tilde{y}, \tilde{X}(\hat{\delta})] = c_{\hat{\omega}}[x(\hat{\delta}), \tilde{X}(\hat{\delta})]\{K_{\hat{\theta}}[\tilde{X}(\hat{\delta})]\}^{-1}\, \tilde{y}$$

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, (δ_i, θ_i), the resulting "ensemble" prediction is generally considered to be more robust. This is in fact the essence of Bayesian Model Averaging.

So rather than using (49), we now construct BMA predictions of *y* [as first proposed by Raftery et al. (1997)]. To do so, recall first (from Section 1.2) 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 $x_l = (x_{l1},..,x_{lk}) \in \mathrm{X}$ 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, (δ_i, θ_i), the corresponding mean prediction for y_l given (δ_i, θ_i) together with data $[\tilde{y}, \tilde{X}]$ is then given by:

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

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

$$(51) \quad E(y_l \mid x_l, \tilde{y}, \tilde{X}) = \frac{1}{N}\sum_{i=1}^{N} E[y_l \mid x_l, \tilde{y}, \tilde{X}, \delta_i, \theta_i]$$

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

$$(52) \quad E(\tilde{y}_j \mid \tilde{x}_j, \tilde{y}, \tilde{X}) = \frac{1}{N}\sum_{i=1}^{N} E[\tilde{y}_j \mid \tilde{x}_j, \tilde{y}, \tilde{X}, \delta_i, \theta_i]$$

$$= \frac{1}{N}\sum_{i=1}^{N} c_{\omega_i}[\tilde{x}_j(\delta_i), \tilde{X}(\delta_i)]\{K_{\theta_i}[\tilde{X}(\delta_i)]\}^{-1}\, \tilde{y}$$

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, (δ_i, θ_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, δ_i.

One approach is simply to ignore such cases by only calculating marginal effects for each explanatory variable, x_{lj}, in those models, δ_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 *continuous* variables, x_{lj}, these are simply taken to be the partial derivatives of mean predictions in (50) with respect to x_{lj}. For *discrete* variables (such as "number of bedrooms" in the housing example below), such partial derivatives should in principle be replaced by appropriate partial differences. But since we wish to compare our results with *Geographical Weighted Regression* (GWR), which also produces local marginal effects, we choose to treat such variables as continuous in order to obtain results more comparable with local regression coefficients. Note finally that explanatory variables may in some cases be *nominal* (such as the "school district" indicator variable in the housing example). While one can in principle evaluate (50) at each alternative nominal value in such cases, we choose here to focus only on quantitative variables.

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

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

For the case of $\delta_{ij}=1$, we may use (50) to obtain the following more explicit form:[6]

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

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

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

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

$$(56)\qquad \frac{\partial}{\partial x_j}c_{\omega_i}[x_l(\delta_i),\tilde{X}(\delta_i)]\ =\ [\frac{\partial}{\partial x_j}c_{\omega_i}(x_{li},\tilde{x}_{i1}),...,\frac{\partial}{\partial x_j}c_{\omega_i}(x_{li},\tilde{x}_{in})]$$

---

[6] 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 paper.

$$= -\tfrac{1}{\tau_i^2}[c_{\omega_i}(x_{li},\tilde{x}_{i1})(x_{lij}-\tilde{x}_{i1j}),..,c_{\omega_i}(x_{li},\tilde{x}_{in})(x_{lij}-\tilde{x}_{inj})]$$

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

$$(57) \qquad ME_{lj} = \frac{1}{N}\sum_{i=1}^{N} ME_{lj}(\delta_i,\theta_i)$$

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

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

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 series of systematic comparisons of this approach with the alternative approaches outlined in the Introduction. We begin in the next section with the simplest of these comparisons, focusing on the BMA versions of spatial regression models proposed by LeSage and Parent [LP] (2007).

3. SIMULATION 1: A Simple Nonseparable Example

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 Section 3.1 with a brief description of these spatial regression models. This is followed in Section 3.2 with a specification of the simulation models to be used, together with comparative simulation results focusing on both model and variable selection.

3.1 SAR-BMA and SEM-BMA Models

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

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

where $X=[x_i : i=1,..,k]$ is an $n\times k$ matrix of explanatory variables (as in Section 2.1 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:

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

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

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

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

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

Note that (unlike [LP]) we use the same symbol, ρ, 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 Bayesian Model Averaging, [LP] follow many of the general conventions proposed by Fernandez, Ley, and Steel [FLS] (2001b). First of all (in a manner similar to our δ vectors in Section 2.2.1 above), if the relevant class of candidate models, *M*, is denoted by M, then each model, $M \in \mathrm{M}$, is specified by a selection of variables (columns) from X, denoted here by X_M, with corresponding parameter vector, β_M. The parameters α and σ together with the relevant spatial autocorrelation parameter, ρ, are assumed to be common to all models, and are given standard non-informative priors. In particular a uniform prior on [-1,1] is adopted for ρ in all simulations below. Only the priors on β_M for each model *M* warrant further discussion, since they utilize data information from X_M. In particular, the prior on β_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 [FLS]) 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 β_M and to avoid sensitivity to units in the associated covariance matrix. For SEM-BMA models, the prior on β_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 β_M, and can thus be said to yield natural "empirical Bayes" priors for β_M.

Aside from the specification of priors, the other key difference between the implementation of SAR-BMA and SEM-BMA in [LP] and our implementation of GPR-BMA in Section 2.2.1 above is the method of estimating both model probabilities and inclusion probabilities. Rather than appeal to asymptotic MCMC frequency approximations as we have done, [LP] 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 Section 2.2 above, and let $p(M \mid \tilde{y}, \tilde{X})$ denote the posterior marginal probability of model M given $(\tilde{y}, \tilde{X})$, then the corresponding *estimated model probabilities*, $\hat{p}(M \mid \tilde{y}, \tilde{X})$, are taken to be these numerical-integration approximations. If for each variable, j, we also let 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

$$(63) \qquad \hat{p}(j \mid \tilde{y}, \tilde{X}) = \sum_{M \in \mathrm{M}_j} \hat{p}(M \mid \tilde{y}, \tilde{X})$$

As verified by [FLS], 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.

3.2 Simulated Model Comparisons

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

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

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

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

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:

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

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

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 the 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. With this in mind, it should be clear that any attempt to approximate such nonseparabilities by appropriate choices of (constant) coefficients, β, 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 insignificant*. This is in fact the 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, β_4 and β_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.

3.3 Simulation Results

The simulation results are displayed in Table 1 below, where the three explanatory variables (x_1, x_2, x_3) and irrelevant variables (z_1, z_2, z_3) are listed in the first column, and where each subsequent column lists the estimated inclusion probabilities (in percentage terms) for the relevant combinations of models and methods.

Table 1

The inclusion probabilities for SAR-BMA and SEM-BMA are based on expression (63) above, and those for GPR-BMA are based on expression (48). The first two columns include results for the SAR-BMA and SEM-BMA models in the separable case described by expressions (64) and (65), respectively. These are included to show that when the true model is linearly separable in (x_1, x_2, x_3), both SAR-BMA and SEM-BMA do extremely well in identifying the correct variables. [The basic SAR and SEM models without BMA (not shown) also do extremely well in terms of standard p-values.] But the most interesting results for our purposes are the last four columns involving the nonseparable case. Here columns three and five compare SAR-BMA and GPR-BMA with respect to the nonseparable specification in expression (66), and similarly, columns four and six compare SEM-BMA and GPR-BMA with respect to the nonseparable specification in expression (67). In both cases it is clear that that even when adding Bayesian Model Averaging to SAR or SEM, these models show little ability to identify the true variables in the presence of such nonseparabilities. Not only are all true *x*-variables unidentified, but they are in fact often less relevant than some of the *z*-variables. This is made even clearer by Table 2 below, where the five models in M with highest estimated marginal posterior probabilities are also shown.

Table 2

Notice that in both cases, the most probable model is precisely the one which includes *no explanatory variables*, i.e., that relies only on the intercept for "explanation". In other words, these globally separable specifications can have difficulty in distinguishing such nonseparable relations from random noise. This is also seen by comparing the overall dispersion of posterior model probabilities. In contrast to the separable case (not shown) where the top five models for both SAR-BMA and SEM-BMA include more than 99.5% of the posterior mass, the situation is quite different in the nonseparable case. For SAR-BMA the top five models only account for 84.8% of posterior mass, and for SEM-BMA it is even less (75.8%). So there is seen to be far more dispersion among alternative candidate models in this nonseparable case. But again we should emphasize that these examples are specifically designed to be challenging for standard linearly specified estimation models, even when spatial autocorrelation components are specified correctly. What is somewhat more surprising is that the extension of these models to include Bayesian Model Averaging seems to offer little in the way of help.

In this light, the single most important result of this simulated comparison is to show how well GPR-BMA is doing with respect to the same data. Even though there is no attempt to capture spatial autocorrelation structure, the true variables are identified 100% of the time, and the irrelevant variables are never identified. 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. It should also be added that Bayesian Model Averaging 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 M. We shall see this again in our second more elaborate simulation example, to which we now turn.

4. SIMULATION 2: A Stylized Housing Price Example

While the above simulation model served to illustrate the consequences of nonseparabilities in a simple and transparent way, there was no attempt to relate this to actual spatial behavior. In the present section we introduce a stylized model of housing price formation in a "Circular City" where nonseparabilities arise from the differences in housing preferences among household types. This example is developed in far more detail, and is used to illustrate the full range of analytical questions that can be addressed with GPR-BMA models. In particular, we consider not only variable selection, but also prediction of unobserved prices, and estimation of the local marginal effects of each variable on prices. In doing so, it is important to compare GPR-BMA with the alternative kernel-based family of Locally Weighted Regression (LWR) models that are specifically designed to estimate local marginal effects. As mentioned in the Introduction, we focus here on Geographically Weighted Regression (GWR), which is by far the most commonly used method for spatial applications. Hence we begin in Section 4.1 below with a brief summary of this method, together with key references where further details can be found. The Circular City Model is then developed in Section 4.2. This is followed in Section 4.3 with a presentation and discussion of the simulation results comparing these two methods.

4.1 Geographically Weighted Regression

Geographically Weighted Regression (GWR) appears to have been introduced independently by McMillen (1996) and Brunsdon, Fotheringham, and Charlton (1996) [but is given its name in the latter paper]. As mentioned in the introduction, our present application of GWR relies more heavily on the approach of McMillen (1996), following Cleveland and Devlin (1988). For given spatial data, $(y_i, x_{i1},..,x_{ik})$, at locations, $i = 1,..,n$, this approach starts with a linear model of the form

$$(68) \qquad y_i = \beta_{i0} + \sum_{v=1}^{k} x_{iv}\beta_{iv} + \varepsilon_i \;=\; x_i'\beta_i + \varepsilon_i \quad , \qquad i = 1,..,n$$

where $x_i = (1, x_{i1},..,x_{ik})'$, $(\varepsilon_1,..,\varepsilon_n) \sim N(0, \sigma^2 I_n)$,[7] and where the coefficient vector, $\beta_i = (\beta_{i0}, \beta_{i1},..,\beta_{iK})'$, is allowed to *vary* across spatial locations $i = 1,..,n$. While (68) appears to be a linear "parametric" model, these spatially varying coefficients can essentially capture any function of spatial locations, so that space itself is implicitly treated nonparametrically.[8] But to estimate this host of parameters, additional

[7] Depending on the context, some developments of GWR make no appeal to independence of residuals. But when formal testing procedures are used [such as in McMillen (1996) and Brunsdon, Fotheringham, Charlton (1999)], this independence assumption is essential. Note also that an extension of this model to include spatial heteroscedasticity is developed in Páez, Uchida and Miyamoto (2002).

[8] In more general LWR models, this same scheme can be employed to model any subset of explanatory variables nonparametrically, as shown in [MR]. Here it should also be noted that the application in [MR]

assumptions are needed. The key assumption here is that parameter variation is sufficiently smooth over space to ensure that parameter values near location i are not "too different" from those at i. This allows parameters, β_i, to be estimated by minimizing weighted sums of squares of the form

(69) $\min_{\beta_i} \sum_{j=1}^{n} (y_j - x_j'\beta_i)^2 w_i(j)$

where parameters at all locations are now replaced by β_i, and where the *local kernel* weights, $w_i(j)$, are implicitly chosen to be close to zero except for those neighboring locations, j, where $\beta_j \approx \beta_i$. This is easily seen to yield a series of standard *weighted least squares* estimates

(70) $\hat{\beta}_i = (X'W_iX)^{-1}X'W_iy \quad , \quad i = 1,..,n$

where $y = (y_1,..,y_n)'$, $X' = (x_1,..,x_n)$, and where W_i is a diagonal matrix with components, $w_i(j)$, $j = 1,..,n$. As with the weight matrices, W, in Section 3.1 above, one must of course pre-specify the local kernel weights, $w_i(j)$. While many choices are possible, we employ the standard *squared exponential* (*Gaussian*) kernel,

(71) $w_i(j) = \exp(-d_{ij}^2 / b)$

which is seen to parallel our choice of the squared exponential covariance kernel in (2) above. Here, d_{ij} denotes Euclidean distance between spatial locations i and j, and parameter $b > 0$ is usually designated as the *bandwidth* of the kernel. Clearly, larger bandwidths include a wider range of relevant locations in (69), and are thus appropriate when spatial variation in β_i coefficients is more gradual. The choice of an appropriate bandwidth is typically carried out by standard "leave-one-out" cross-validation techniques [as for example in Brunsdon, Fotheringham, and Charlton (1996, Section 3.2)]. Given this basic GWR model, we now focus on procedures for prediction, variable selection and local marginal analysis within this framework.

4.1.1 Prediction in GWR

Here we simply sketch the basic implementation of prediction in GWR, and refer the reader to Harris, Fotheringham, Crespo, Charlton (2010) for further details. To begin with, notice that while the basic model in (67) was developed only at data points (to provide a natural comparison with standard OLS), it should be clear that for any *target location*, s, where attribute data, $x_s = (1, x_{s1},..,x_{sk})'$, is available, (68) can be extended to include location *s* as:

(72) $y_s = \beta_{s0} + \sum_{v=1}^{k} x_{sv}\beta_{sv} + \varepsilon_s = x_s'\beta_s + \varepsilon_s \quad , \quad \varepsilon_s \sim N(0, \sigma^2)$

---

employs an alternative version of LWR, namely Conditional Parametric Regression (CPAR), which differs from GWR by including spatial coordinates among the explanatory variables of the model.

So even though y_s is not observed, it can be predicted by estimating the conditional expectation

(73) $E(y_s \mid x_s) = x_s'\beta_s$

To do so, one can estimate β_s by a corresponding extension of (70). In particular, by letting d_{is} denote the Euclidean distance from i to s, one can extend $w_i(j)$ in (71) to a *local kernel function*, $w_i(s)$, on the entire space of locations. If the corresponding kernel matrix at *s* is denoted by, $W_s = diag[w_i(s) : i = 1,..,n]$, then β_s can be estimated as

(74) $\hat{\beta}_s = (X'W_sX)^{-1}X'W_sy$,

This in turn yields the following estimate of $E(y_s \mid x_s)$,

(75) $\hat{y}_s = \hat{E}(y_s \mid x_s) = x_s'\hat{\beta}_s = x_s'(X'W_sX)^{-1}X'W_sy$,

which provides a mean prediction of y_s based on the observed data. For example, in our housing price model below, if y_i is the observed sales price of a house with attributes, x_i, at locations $i = 1,..,n$, and if the attributes, x_s, of some unsold house at location *s* are available, then (75) yields a natural prediction of the sales price, y_s, for this house based on observed prices of its neighbors. Notice finally that there is some degree of parallel between the predictions in (75) and GPR predictions in expression (8), where the covariance kernel, $c_\omega(\cdot)$, is replaced by the local kernels, $w_s(\cdot)$. However, the key difference here is that $w_s(\cdot)$ involves only 2-dimensional distances between spatial locations, while $c_\omega(\cdot)$ involves *k*-dimensional distances between all explanatory variables.

4.1.2 Variable Selection in GWR

While there are in principle many ways to carry out variable selection within this GWR framework, the simplest and most intuitive approach (in our view) is the semiparametric method of McMillen and Redfearn (2010) [MR] mentioned in the Introduction. If for any candidate explanatory variable, v, we denote the set of all other explanatory variables by $\tilde{v} = \{1,...,k\} - v$, then the essential idea (in the present setting) is to use GWR to "remove" the influence of $\tilde{v}$ on both y and *v*, and then run a simple regression using these residuals to determine the relevance of *v* on *y* in the absence of $\tilde{v}$.[9] To implement this procedure, we first eliminate variable v by setting $x_i(\tilde{v}) = (x_{ij} : j \in \tilde{v})$ and $X(\tilde{v})' = [x_i(\tilde{v}) : i = 1,..,n]$. The GWR prediction of y_i based on variables, $\tilde{v}$, is then given by

[9] This semiparametric procedure is conceptually very similar to the "mixed" GWR model developed in Chapter 3 in Fotheringham, Brunsdon, and Charlton (2002). In that setting, the present candidate variable, x_v, would be formally treated as the "parametric" part of the model.

(76) $\hat{y}_i(\tilde{v}) = x_i(\tilde{v})'[X(\tilde{v})'W_iX(\tilde{v})]^{-1}X(\tilde{v})'W_iy$

and similarly, the GWR prediction of x_{vi} based on $\tilde{v}$ is[10]

(77) $\hat{x}_{vi}(\tilde{v}) = x_i(\tilde{v})'[X(\tilde{v})'W_iX(\tilde{v})]^{-1}X(\tilde{v})'W_ix_{vi}$

Given these predictions, if we denote the portion of y_i not explained by $\tilde{v}$ as

(78) $\Delta y_i(\tilde{v}) = y_i - \hat{y}_i(\tilde{v})$

and the portion of x_{vi} not explained by $\tilde{v}$ as

(79) $\Delta x_{vi}(\tilde{v}) = x_{vi} - \hat{x}_{vi}(\tilde{v})$,

then the explanation of *y* provided by variable v independent of all other variables, $\tilde{v}$, can be gauged by the results of the simple regression:

(80) $\Delta y_i(\tilde{v}) = \beta_0 + \beta_v\Delta x_{vi}(\tilde{v}) + \varepsilon_i \quad , \quad i = 1,..,n$

Following [MR], it is most natural to use the *p-value* of β_v to measure the statistical significance of variable v in providing additional information about y. Note finally that (unlike the inclusion probabilities in Section 2.2.2 above) the *sign* of β_v provides some information about the overall direction of influence on *y*.

4.1.3 Local Marginal Analysis in GWR

In many ways, local marginal analysis is the simplest of these procedures in GWR, since this was the original purpose of the model itself. As a parallel to (54) above, it now follows from (73) that for any explanatory variable, v,

(81) $\frac{\partial}{\partial x_v}E(y_s \mid x_s) = \frac{\partial}{\partial x_v}(x_s'\beta_s) = \beta_{sv}$

so that estimates of such local marginal effects are given precisely by the estimated local beta coefficients, $\hat{\beta}_{sv}$. In principle, one can thus use the standard errors from each local regression to obtain confidence bounds on the magnitude of β_{sv}. But as pointed out by Wheeler and Tiefelsdorf (2005), GWR is much more sensitive to multicollinearity among explanatory variables than is ordinary regression.[11] So great care must be taken in interpreting these values, especially for small sample sizes.[12]

---

[10] Here it should be noted that cross-validation bandwidths could in principle differ between (75) and (76). But for consistency we use the same bandwidth in (76) for each variable in (77). So the W_i matrices are in fact the same, as shown.

[11] Wheeler and Tiefelsdorf (2005) propose checking standard diagnostics like Variance Inflation Factors. Visual diagnostic methods are also developed in Wheeler (2010).

[12] The detailed simulation study by Páez, Farber, and Wheeler (2011) suggests that estimates of individual beta coefficients tend to be unreliable for sample sizes less than $n \approx 160$.

4.2 The Circular City Model

In [MR] a stylized one-dimensional model of urban population density was developed to illustrate the superiority of LWR over OLS in fitting spatially varying functions. In the present section we develop a two-dimensional stylized model of urban housing prices which is more explicit in terms of explanatory variables, and allows a wider range of estimation issues to be addressed, including both variable selection and local marginal analysis. In particular, we consider a small circular city of one-mile radius with three inner zones (Z_1, Z_2, Z_3) and three outer zones (Z_4, Z_5, Z_6) surrounding the CBD, as shown in the left panel of Figure 2 below. Each zone is assumed to contain approximately 500 residential parcels, yielding a total of about 3000 parcels for the entire city. In each zone, these parcels are distributed on a uniform grid (so that parcels in the smaller inner zones are more densely distributed).

Figure 2

The local population is assumed to consist of two types of households. First there are about 2000 young families with children and with bread-winners working in the CBD (which is here idealized to be a point location). These families are naturally interested in homes with more bedroom space and located closer to the CBD. In addition, it is assumed that all children are of school age, so that proximity to schools is important for each of these families. As shown in Figure 2, there are two school districts, with District 1 consisting of zones (Z_1, Z_4) and District 2 consisting of zones (Z_2, Z_5). School busing is provided within each district, so that there is a strong incentive for families to live in one of these two districts. The remaining households (about 1000 in number) are assumed to consist mostly of older retired couples (or individuals) who have smaller space needs, and no need for either school busing or commuting to the CBD. Consequently, their preferences are quite different from the younger families.

In this framework, housing prices are assumed to be generated by competitive bidding. In particular, it is assumed that families are strongly motivated to outbid retirees for properties in school districts, so that housing prices in Districts 1 and 2, consisting of zones (Z_1, Z_2, Z_4, Z_5), reflect the preferences of these families. Similarly, housing prices in zones (Z_3, Z_6) are assumed to reflect the preferences of retirees. To model the price per square foot, P_i, for each house, i, we now let B_i denote the number of bedrooms in i, and let D_i denote the (straight line) distance from i to the CBD. Finally, letting the indicator variable, S_i, denote whether i is in a school district or not, the expected price, $E(P_i)$, for house i is assumed to be of the following (highly nonlinear) form:

$$(82) \qquad E(P_i) = a_0 S_i \left[\exp\{a_1 B_i^{b_1} - a_2 (1+D_i)^{d_1}\} \right]^{a_3} + |S_i - 1| \left[a_4 \{(1+D_i)^{d_2} / B_i^{b_2}\} \right]$$

where all parameters $(a_0, a_1, a_2, a_3, a_4, b_1, b_2, d_1, d_2)$ are positive.[13] To interpret this expression, note first that by construction, $E(P_i)$ is equal to the first term for those houses in school districts ($S_i = 1$), and is equal to the second term for houses outside school districts ($S_i = 0$). Hence housing prices in school districts are seen to be increasing in the number of bedrooms and decreasing in distance from the CBD. On the other hand, these relations are reversed for houses not in school districts. Here retirees are assumed to prefer smaller residences that are more easily maintained. Moreover, they are not working and are thus assumed to prefer living away from the noise and traffic of the CBD.

In this setting, the existing housing stock in each zone is assumed to have average values for each relevant variable as shown in Table 3 (where the bottom row can be ignored for the present).

Table 3

Values of *S* and average values of *D* are direct consequences of the city layout in Figure 2. With respect to bedrooms, *B*, note that these average values do not perfectly match the desires of relevant households in each zone, and thus that there must be tradeoffs between these attributes. But the single most important feature of these values for our purposes is the *nonseparability* of mean prices, $E(P)$, created by the different preferences of consumer types [in a manner paralleling the simpler examples in expressions (66) and (67) above]. The average values of these mean prices are given in the last row of Table 3, and a more detailed representation of the overall spatial price pattern is shown in the right panel of Figure 2. Here the highest expected prices are in Zones 1 and 6, which contain the most preferred house-location combinations for young families and retirees, respectively. (Note also that the sharp contrasts in housing values near the center are mainly a consequence of our simplifying assumption that the CBD is a single point.)

4.3 Simulation Results

Using this second spatial simulation, we investigate how well each technique does on the twin tasks of explanation and prediction by examining, in turn, their ability to correctly classify the statistical relevance of individual explanatory variables, and their ability to predict out-of-sample price values. We first present the variable selection results for all four techniques, and then discuss out-of-sample performance for GPR-BMA and GWR. These metrics are evaluated using 10 different simulations per sample size with sample sizes ranging from 60 (10 observations per candidate explanatory variable) to 270 (45 observations per candidate explanatory variable) by increments of 30.[14]

---

[13] The specific parameter values chosen for simulation purposes are $a_0 = 104.75$, $a_1 = 1.041$, $a_2 = 1.0685$, $a_3 = 0.1792$, $a_4 = 54.598$, $b_1 = 0.9328$, $b_2 = 0.1083$, $d_1 = 1.9334$, and $d_2 = 1.5186$.

[14] Other implementation details include the following. All parameter and model vectors are tested for convergence after each 10 iterations, and a jitter of 0.01 is used for numerical stability. Whenever the model vector changes in the outer loop, the inner loop HMC procedure uses 30 iterations. Otherwise, only 5 additional iterations are used. In addition, this HMC procedure takes 10 steps in each iteration, with step adjustments of 0.05. Finally, the first 75 passes of the full Gibbs sampling procedure are used as burn-in.

4.3.1 Results for Variable Selection

The assessment of variable selection ability proceeds in a fashion similar to Simulation 1. We augment the three *model variables*, (D,B,S), with three *non-model variables*, (z_1,z_2,z_3), each independently normally distributed with individual spatial correlation. Our criterion for an estimation method to be successful is that it should identify (D,B,S) as statistically relevant and (z_1,z_2,z_3) as statistically irrelevant. Given that these four methods include both Bayesian and non-Bayesian approaches, we use two different measures of statistical relevance.

For Bayesian methods (SAR-BMA, SEM-BMA, GPR-BMA) statistical relevance is evaluated in terms of posterior variable-inclusion probabilities [as in expressions (48) and (63) above]. Here higher inclusion probabilities are taken to imply stronger relevance for individual variables. For the non-Bayesian method of GWR, statistical relevance is evaluated in terms of p-value calculations for its associated semiparametric test [following expression (80) above]. It should be noted that p-value scaling is the opposite of inclusion probabilities, with lower p-values denoting higher levels of statistical significance. In this context, our specific evaluation criteria are that the model variables, (D,B,S), should all have inclusion probabilities of no less than 0.95 (or p-values of no greater than 0.05). Similarly, the non-model variables, (z_1,z_2,z_3), should have inclusion probabilities less than 0.95 (or p-values greater than 0.05).

Table 4 contains four panels, one for each estimation method. Panels A, B, and C show the results for the Bayesian methods, SAR-BMA, SEM-BMA, and GPR-BMA, respectively. In particular, Panels A through C contain inclusion probabilities averaged across 10 simulation runs for each of the sample sizes shown. Similarly, Panel D contains p-values for GWR averaged across these 10 simulations.

Table 4

The SAR-BMA results (Panel A) and SEM-BMA results (Panel B) are seen to be similar to those in Simulation 1. Again as a result of their separable parametric specifications, neither method is able to identify any of these strongly nonseparable variables, (D,B,S), as being statistically relevant (at any sample size). In fact, these inclusion probabilities are not substantially higher than those for non-model variables, with the maximum value, 0.26, being far less than the 0.95 needed for statistical relevance. Moreover, no noticeable improvement occurs with increasing sample size.

Turning next to the GWR results in Panel D, we see substantial improvement with respect to model variables *B* and *S*, which are both strongly significant for all sample sizes of 90 and higher, while all non-model variables are very insignificant. The only notable failure is the inability of GWR to identify distance, *D*, as statistically significant. More generally, this inability to identify variables that are systematically dependent on

location appears to be an inherent property of locally weighted regression methods, as discussed further by [MR].[15]

Finally we turn to the GPR-BMA results in Panel C. As in Simulation 1, this Bayesian nonparametric method achieves essentially perfect identification of both model and non-model variables. These strong results appear to be attributable to the fact that regardless of the random initialization of the model vector, δ, GPR-BMA quickly converges to the true model. However, it should be emphasized that this degree of accuracy occurs at a price. In particular, a comparison of time scales for GWR and GPR-BMA shows that the latter is relatively very costly in terms of computation time. So the present version of this model is limited to small and medium sized data sets. This limitation is currently a very active area of research, as discussed further in the concluding section of the paper.

4.3.2 Results for Out-of-Sample Prediction

Turning next to out-of-sample predictions, we shall focus exclusively on the two nonparametric estimation methods, GWR and GPR-BMA. The flexibility of these methods with respect to unknown relationships between the dependent and independent variables makes them particularly well suited for out-of-sample predictions. As noted by McMillen et al. (2010, 2012), this is particularly true in spatial contexts, where such approaches avoid many of the misspecification problems inherent, for example, in standard spatial lag models.

In addition, while SAR-BMA and SEM-BMA provide only global marginal effects, GWR and GPR-BMA yield localized out-of-sample predictions for both values of the dependent variable and marginal effects of the explanatory variables. The relevant marginal effects in the present simulation model involve both the effects of distance (ME-D) and number of bedrooms (ME-B). With respect to ME-B in particular, the number of bedrooms, *B*, is treated as a smooth variable in GPR-BMA to maintain some degree of comparability with the regression-slope estimates in GWR. In addition, GWR predictions are based on the *true* set of explanatory variables rather than on the statistically relevant variables (since *D* would be excluded). In contrast, predictions in GPR-BMA are based on the selected model vector and set of parameters. Also, while GWR only has one set of localized forecasts per simulation, GPR-BMA has numerous forecasts based on multiple draws of model vectors and parameters. Consequently, GPR-BMA uses all models from each post burn-in run across the 10 simulations to calculate predictions. Finally, prediction accuracy is here measured in terms of root mean square error (RMSE), where lower values of RMSE are taken to imply more accuracy. The results of these calculations are shown in Table 5 below:

Table 5

[15] [MR] suggest that the optimal bandwidth parameter may depend on the type of variable. In particular, they recommend [following Pagan and Ullah (1999)] that larger window sizes should be used for locational variables such as distance. How much larger still remains an important and unanswered question. In unreported results, we doubled the size of the optimal bandwidth parameter, yet did not find any change in the statistical relevance of the distance variable. (In contrast to the semiparametric test, GPR-BMA uses the same set of three parameters regardless of variable type.)

For both techniques, price prediction is seen to improve (lower RMSE) with increased sample size. While the performance of GWR improves at a faster rate, the corresponding results for GPR-BMA are uniformly sharper. In particular, the RMSE values for GWR are everywhere more than twice as high as those for GPR-BMA (and often three times as high).[16]

The results for marginal effects are very similar. GPR-BMA provides more accurate predictions of marginal effects at every sample size. In relative terms, the performance of GWR is noticeably worse than for price prediction, with RMSE values ranging from 2.7 to 4.8 times higher than those for GPR-BMA.[17] But again, the area where GWR does excel is in terms of computation time, which is nearly an order of magnitude better than GPR-BMA.

We next observe that the marginal-effect results in Table 5 deal exclusively with error magnitudes. But often the most important question about such effects is their direction. For example, since the critical nonseparabilities in the present model are generated precisely by sign reversals in preferences among household types, it is of particular interest to determine how well these reversals are picked up by each estimation method.

Before doing so, however, it is important to recognize one limitation inherent in such local analyses of marginal effects. Even when sample sizes are large and spatially dense, there is always a problem of *multiple testing* that arises from the overlap of regression neighborhoods. In particular, this implies that marginal estimates at nearby locations must necessarily be *positively correlated*, which makes it more difficult to gauge the joint significance of marginal effects between such locations. While these dependencies are mitigated to a certain degree by model-averaging procedures such as GPR-BMA, they are still present. Moreover, while there have been numerous efforts to discount such effects in a systematic way, this issue remains an ongoing area of active research. [For a discussion these issues in the specific context of GWR see Charlton and Fotheringham (2009). For more general reviews of such work in a spatial context see Castro et al. (2006) and Robertson et al. (2010).] So the approach adopted here is simply to employ standard diagnostic methods, and to compare the results of these diagnostics for both GWR and GPR-BMA. One advantage of the present simulation framework is that such diagnostics can be directly compared against true values to see how well they perform in the presence of such correlations.

---

[16] Here it should also be noted that the prediction accuracy of GWR has been compared with non-Bayesian GPR (Kriging) by Harris et al. (2010, 2011). While these comparisons focus more on non-stationary versions of these models, it is nonetheless clear that in terms of prediction accuracy their results are qualitatively similar to ours.

[17] It should be noted however the bandwidth parameter for GWR is optimized only with respect to prediction of the dependent variable, though it is used also for marginal effects. This may in part account for the slight degradation in performance with respect to marginal effects. It is also of interest to note that [MR] have suggested (in addition to footnote 12 above) that appropriate bandwidths for marginal effects should probably be larger than those for value predictions. But this question has not been pursued further in the present paper.

There is one final issue that must be addressed in comparing Bayesian versus non-Bayesian methods, namely the absence of a classical testing framework in Bayesian approaches that is based on "null hypotheses". But since there is some degree of comparability between Bayesian *credible intervals* and non-Bayesian *confidence intervals*, we choose this approach for purposes of comparison.[18] In this context, our basic diagnostic approach is to construct appropriate one-sided 95% confidence (or credible) intervals to gauge the "credibility" of signs. In the case of GWR, the appropriate confidence intervals are for the coefficients, β_{sv}, of variables, v, in the (weighted) regression at each location s [in expression (81) above]. In particular, β_{sv} is said to be *significantly positive* (*negative*) if the upper (lower) 95% confidence interval for β_{sv} lies above zero (below zero). [For any threshold value, β_0, such upper and lower 95% confidence intervals for β_{sv} are of the form, $[\beta_0, \infty)$ and $(-\infty, \beta_0]$, respectively.] Moreover, since it is well known (from the symmetry of normal and *t*-distributions) that the critical region for an upper (or lower) tailed test of β_{sv} at the 0.05 level is precisely the upper (or lower) 95% confidence interval for β_{sv}, it follows that such intervals can be constructed entirely in terms of the standard *t*-statistics for $\hat{\beta}_{sv}$. In the case of GPR-BMA, such beta parameters, β_{sv}, are replaced by the BMA *marginal effect*, ME_{sv}, of explanatory variable, *v*, at location, *s* [where for sake of comparison we here replace *l* and *j* in expression (57) above with *s* and *v*, respectively]. Here the appropriate credible intervals are based on the posterior distribution of ME_{sv}. In particular, the marginal effect of variable *v* and location *s* is said to be *credibly positive* (*negative*) if the upper (lower) 95% credible interval for ME_{sv} lies above zero (below zero) [where for any threshold value, ME_0, the upper and lower credibility intervals for ME_{sv} have the same respective forms, $[ME_0, \infty)$ and $(-\infty, ME_0]$, as above]. These credible intervals are estimated from the corresponding frequency distribution of ME_{sv} obtained from Gibbs sampling (which approximates the posterior distribution of ME_{sv}).

Before reporting these comparative results, we note finally that a distinction must be made between "significant effects" and "correct effects". While it is of course desirable that an estimation method identifies the signs correctly in those cases deemed to be significant, it is also important that a large fraction of cases with true non-zero signs actually be deemed significant. With this in mind, we report both measures. In particular, the fractions of those (out-of-sample) cases deemed to be significant (either positive or negative) for each method and the fractions of these cases with correct signs are both shown in Table 6 below. Turning to the left panel of Table 6, we see that both GPR-BMA and GWR do quite well at identifying non-zero marginal effects as significant.

Table 6

[18] There do exist Bayesian variants of GWR [Lesage (2004)] that would allow Bayesian credible intervals to be used for both GWR and GPR-BMA. But since such Bayesian versions of GWR are used far less frequently, we opt for the standard non-Bayesian approach.

Notice also that for both methods, the fractions of cases deemed significant for ME-D are uniformly higher than for ME-B. This is not surprising in view of the fact that "marginal effects" themselves are more problematic for the discrete number-of-bedrooms variable, *B*, then for the continuous distance variable, *D*.

This difference in results between ME-D and ME-B is even more evident when comparing the relative performance of these two methods. For the continuous variable, *D*, the results for GPR-BMA are uniformly higher than for GWR (in a manner similar to, but less dramatic than, the prediction results above). However, for the discrete variable, *B*, it appears that while GPR-BMA performs better for small samples, the results are mixed for larger samples. Here, in addition to the general "discreteness" problem mentioned above, it should be noted that in this particular simulation model most of the "insignificant" results for both methods occur in Sector Z_3 of Figure 2 close to the CBD, where (as can be seen from the sharp price variations in the right panel of Figure 2) values of *B* tend to exhibit variations that are exaggerated by the idealized assumption of a point CBD. So the "smoothness" assumptions implicit in such marginal analyses are most severely tested in this region.

Turning finally to the fractions of correct signs occurring in cases deemed to be significant as shown in the right panel of Table 6, the overall pattern seems roughly similar to the previous set of results. But there are two key differences that should be stressed. First, GWR is performing *uniformly worse* than in the previous exercise. Second, GPR-BMA is performing *uniformly better* than GWR. So at least in this simulated model, the occurrence of *significant but incorrect signs* for local marginal effects appears to be much less of a problem for GPR-BMA than for GWR.

5. EMPIRICAL EXAMPLE: Predictors of Economic Growth

In this final section, we apply GPR-BMA to a real dataset to highlight how well this technique performs in practice. Here we use a standard BMA benchmark dataset focusing on economic growth [FLS (2001), Sala-i-Martin (1997)], which includes 42 candidate explanatory variables for each of 72 countries. To capture possible spatial effects, we include the spatial location (latitude 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 dataset, we provide OLS-BMA results alongside GPR-BMA results. To facilitate this 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 (≈ 10.5) was taken as the mean of the prior distribution for *q* in expression (19) and used to solve numerically for λ, yielding a value of $\lambda = 0.088$ (as shown in Figure 1 above). The resulting *Variable Inclusion Probabilities* (V.I.P.) for both OLS-BMA and GPR-BMA are shown in in the first two columns of Table 7 below

[where the OLS-BMA results were calculated using Matlab code from Koop, Poirier and Tobias (2007)]. The first two rows include the spatial variables, and the remaining rows are ordered by inclusion probabilities under GPR-BMA.

Table 7

Observe first that there is strong agreement between these two sets of inclusion probabilities, with an overall correlation of nearly 85%. In particular, both methods are in agreement as to the most important variables (with inclusion probabilities above .90), with the single exception, *Non-Equipment Investment*. Here the inclusion probability under GPR-BMA (.947) is more than twice that of OLS-BMA. 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 of OLS is well known to be more sensitive to such colinearities than GPR, this could well be the main source of the difference.

Turning next to the spatial variables, latitude and longitude, it is clear that neither is a relevant predictor of economic growth in the present data set. So even though the inclusion probabilities for latitude are uniformly higher than those for longitude, these values provide little support for the Sachs (2001) hypothesis.[19] Note also from the specification of the squared exponential kernel in (2) that the presence of both 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).

5.1 Average Marginal Effects

Estimates of the *Average Marginal Effects* (A.M.E.) of each variable are shown in the last two columns of Table 7, where it is again seen that the results for OLS-BMA and GPR-BMA are quite similar. Moreover, this similarity is even stronger when one considers the influence of inclusion probabilities on marginal effects [as seen for GPR-BMA in expression (53) above]. In particular, differences in average marginal effects between these methods are often the result of corresponding differences between their associated inclusion probabilities. 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 average marginal effect of this variable in GPR-BMA is also roughly twice that in OLS-BMA, one can conclude that average marginal effects restricted to those models where Non-Equipment Investment is present are actually quite similar for these two methods.

[19] A more direct test of this particular hypothesis would be to use 'absolute latitude' (to distinguish between tropical and temperate zones), as is done in both Sala-i-Martin (1997) and FLS (2001). But since experiments with this variable produced lower inclusion probabilities than latitude, we chose to report only results for the latter.

5.2. Local Marginal Effects

But while average marginal effects under GPR-BMA are similar to those under OLS-BMA, it is possible to probe deeper with GPR-BMA. Unlike OLS-BMA, where the marginal effect of each variable is 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 OLS-BMA. As one illustration, we now consider differences between the marginal effects of *Equipment Investment* (*E_Inv*) and *Non-Equipment Investment* (*NE_Inv*) across countries, as displayed in Table 8 below (where the marginal effects of equipment investment are shown in descending order).

Table 8

Not surprisingly, the highest marginal effects of equipment investment are exhibited by less developed countries (such as Cameroon) and the lowest marginal effects by more developed countries (such as the United States). But a more interesting relation can be seen by plotting marginal effects for both E_Inv and NE_Inv against their corresponding investment levels for each country, as shown in Figure 3 (where circles and stars are used to represent marginal effects for E_Inv and NE_Inv, respectively).

Figure 3

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 (-0.76) is indeed much steeper than that for other investments (-0.43) [as seen graphically by the regression lines plotted in Figure 3]. In summary, such results serve to illustrate how GPR-BMA can be used to address a wide range of questions not accessible by the more standard OLS-BMA approach.

6. Concluding Remarks

The objective of this paper 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 Section 3.3 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 GWR. 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). Moreover, the focus of these models on local effects of variables tends to ignore the possible global relations among them.[20] 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 Bayesian Model Averaging. But as we have seen in both SAR-BMA and SEM-BMA, such model averaging by itself may not be very effective when un-modeled 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 τ.

This ability to capture both global and local interactions has a wide range of applications in empirical analyses, as seen in the economic growth example of Section 5. Here we saw that GPR-BMA was not only able to capture global determinants of economic growth in a manner similar to OLS-BMA, but was also able to delve deeper. In particular, the localized marginal effects 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

[20] While there are indeed "mixed" versions of such models that incorporate both global (parametric) and local (nonparametric) specifications [as detailed for example 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.

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 [RW]). 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 et. al. (2005) for an early example]. Solutions range from direct sub-sampling 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 Bayesian Model Averaging 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 Section 5. Such extensions will be reported in a subsequent paper.

References

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

Brunsdon, C., A.S. Fotheringham, and M.C. Charlton (1996) "GeographicallyWeighted Regression," *Geographical Analysis*, 28, 281–298.

Brunsdon, C., A.S. Fotheringham, and M.C. Charlton (1999) "Some notes on parametric significance tests for geographically weighted regression", *Journal of Regional Science*, 39: 497-524.

Castro, M. and Singer, B. (2006) "A new approach to account for multiple and dependent tests in local statistics of spatial association: controlling the false discovery rate", *Geographical Analysis*, 38: 180–208.

Chalupka, K., Williams, C., and I. Murray (2013) "A framework for evaluating approximation methods for Gaussian process regression", *Journal of Machine Learning Research*, 14: 333-350

Charlton, M. & Fotheringham, S. (2009) *Geographically weighted regression*, National Centre for Geocomputation, Maynooth, Ireland.

Chen, T. and B. Wang (2010) "Bayesian variable selection for Gaussian process regression: Application to chemometric calibration of spectrometers", *Neurocomputing*, 73: 2718-2726.

Cleveland, W.S., and S.J. Devlin (1988) "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of American Statistical Association*, 83: 596-610.

Cornford, D., Csató, L., & Opper, M. (2005). Sequential, Bayesian geostatistics: a principled method for large data sets. *Geographical Analysis*, *37*(2), 183-199.

Cotteleer, G., Stobbe, T., and G. van Kooten (2011) "Bayesian model averaging in the context of spatial hedonic pricing: an application to farmland values", *Journal of Regional Science*, 51: 540-557.

Denison, D.G.T., B.K. Mallick, and A.F.M. Smith (1998) "Bayesian MARS", *Statistics and Computing*, 8: 337-346.

Duane, S., Kennedy, A., Pendleton, B., and D. Roweth (1987) "Hybrid Monte Carlo", *Physics Letters B*, 195: 216-222.

Fotheringham, A. S., & Brunsdon, C. (1999). Local forms of spatial analysis. *Geographical Analysis*, *31*(4), 340-358

Fotheringham, A. S., C. Brunsdon, and M. Charlton (2002) *Geographically Weighted Regression: the analysis of spatially varying relationships*, Wiley: New York.

Fernandez, C., Ley, E., and M. Steel (2001a) "Model uncertainty in cross-country growth regressions", *Journal of Applied Econometrics*, 16: 563-576.

Fernandez, C., Ley, E., and M. Steel (2001b) "Benchmark priors for bayesian model averaging", *Journal of Econometrics*, 100: 381-427.

Gabrosek, J., & Cressie, N. (2002). The effect on attribute prediction of location uncertainty in spatial data. *Geographical Analysis*, *34*(3), 262-285

Gahegan, M. (2000). On the application of inductive machine learning tools to geographical analysis. *Geographical Analysis*, *32*(2), 113-139

George, E. I. and R.E. McCulloch (1993) "Variable selection via Gibbs sampling", *Journal of the American Statistical Association*, 88: 881-889.

Getis, A., & Griffith, D. A. (2002). Comparative spatial filtering in regression analysis. *Geographical analysis*, *34*(2), 130-140

Green, P. J. (1995) "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination", *Biometrika*, 82: 711-32.

Harris, P. , A.S. Fotheringham , R. Crespo and M. Charlton (2010) "The Use of Geographically Weighted Regression for Spatial Prediction: An Evaluation of Models Using Simulated Data Sets", *Mathematical Geosciences*, 42: 657–680

Harris, P., C. Brunsdon and A.S. Fotheringham (2011) "Links, comparisons and Extensions of the geographically weighted regression model when used as a spatial predictor", *Stochastic Environmental Research and Risk Assessment*, 25: 123-138.

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*, DOI: 10.1080/ 10618600.2013.812872.

LeSage, J. (2004) "A family of geographically weighted regression models", *Advances in Spatial Econometrics: Methodology, Tools and Applications*, (eds) Anselin, L., Florax, R. and S. Rei, pp. 241-264.

LeSage, J., and O. Parent (2007) "Bayesian model averaging for spatial econometric models", *Geographical Analysis,* 39: 241-267.

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*, Chapman-Hall: Boca Raton, Florida.

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 C.M.Bishop (Ed.), *Neural Networks and Machine Learning*, Springer: Berlin, pp. 133–165.

Matheron, G. (1963) "Principles of geostatistics", *Economic geology*, 58: 1246-1266.

McMillen, D. (1996) "One hundred fifty years of land values in Chicago: A nonparametric approach", *Journal of Urban Economics*, 40:100-124.

McMillen, D. (2012) "Perspectives on spatial econometrics: linear smoothing with structured models", *Journal of Regional Science*, 52: 192-209.

McMillen, D., and C. Redfearn (2010) "Estimation and hypothesis testing for nonparametric hedonic house price functions", *Journal of Regional Science*, 50: 712-733.

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: 587-598.

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

Neal, R. M. (2010) "MCMC using Hamiltonian dynamics", Chapter 5 in *Handbook of Markov Chain Monte Carlo*, eds. S. Brooks, A. Gelman, G. Jones, and X. L. Meng, Chapman and Hall:CRC.

Páez, A., T.Uchida and K. Miyamoto (2002) "A general framework for estimation and inference of geographically weighted regression models: 1. Location-specific kernel bandwidths and a test for locational heterogeneity", *Environment and Planning A*, 34:733–754

Páez, A., Farber, S., and D. Wheeler (2011) "A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships", *Environment and Planning-Part A*, 43: 2992-3010.

Pagan, A. and A. Ullah (1999) *Nonparametric Econometrics*, New York: Cambridge University Press.

Raftery, A., Madigan, D., and J. Hoeting (1997) "Bayesian Model Averaging for Linear Regression Models". *Journal of the American Statistical Association*, 92: 179-191.

Rasmussen, C., and C. Williams (2006). *Gaussian process for machine learning.* MIT Press.

Robertson, C., T. A. Nelson, Y.C. MacNab, and A.B. Lawson (2010) "Review of methods for space–time disease surveillance", *Spatial and Spatio-Temporal Epidemiology*, 1: 105-116

Robinson, P.M. (1988) "Root-N-Consistent Semiparametric Regression", *Econometrica*, 56: 931-954.

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: 111-132.

Sachs, J. (2001) "Tropical underdevelopment", *National Bureau of Economic Research, Working Paper 8119*.

Seeger, M., C.K.I Williams, and N.D. Lawrence (2003) "Fast forward selection to speed up sparse Gaussian process regression", *Workshop on AI and Statistics* 9.

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

Wei, C-H. and F. Qi (2012) "On the estimation and testing of mixed geographically weighted regression model", *Economic Modelling*, 29: 2615–2620

Wheeler, D. and M. Tiefelsdorf (2005) "Multicolinearity and correlation among local regression coefficients in geographically weighted regression", *Journal of Geographical Systems*, 7: 161-187.

Wheeler, D.C. (2010) "Visualizing and diagnosing coefficients from geographically weighted regression'", in *Geospatial Analysis and Modeling of Urban Structure and Dynamics*, Eds. B Jiang, X Yao, Springer: New York, pp 415-436.

Williams, C. K. I. and Rasmussen, C. E. (1996) "Gaussian processes for regression", in Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems 8*, pages 514-520. MIT Press: Boston.

Yan, F. (2010) "Sparse Gaussian process regression via L1 penalization", *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*.

Yi, G., Q. Shi, and T. Choi (2011) "Penalized Gaussian process regression and classification for high-dimensional nonlinear data", *Biometrics*, 67:1285-1294.

Figures

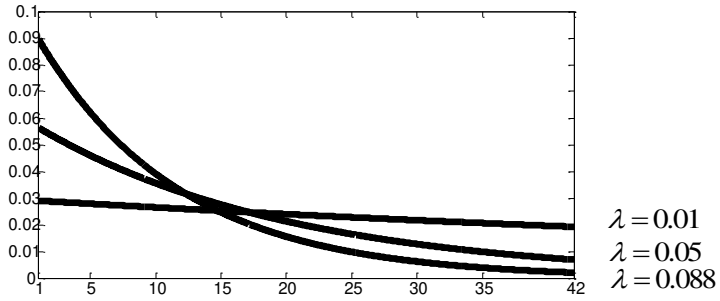

Figure 1. Selected λ values

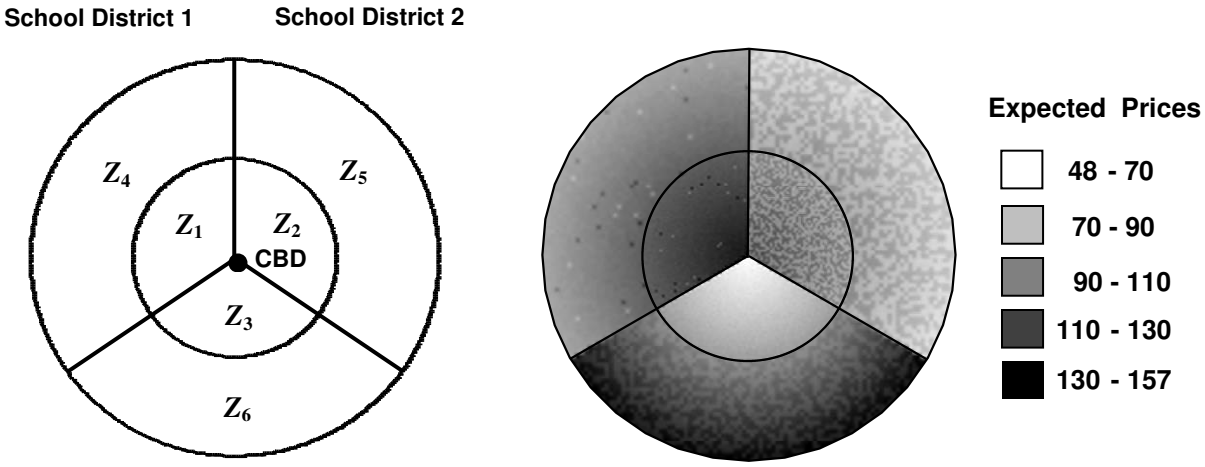

Figure 2. Circular City Layout

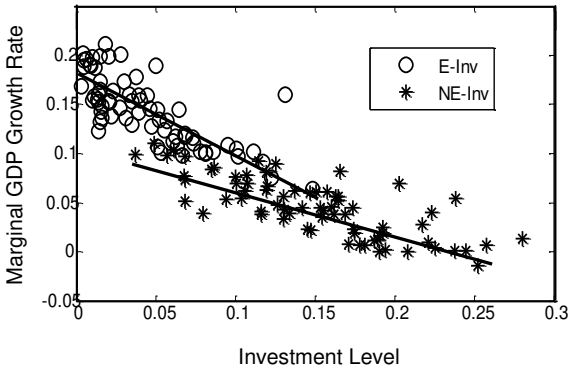

Figure 3. Investment Comparisons

Tables

	SEPARABLE		*NONSEPARABLE*		*NONSEPARABLE*	
Variables	**SAR-BMA Eq. 65**	**SEM-BMA Eq. 66**	**SAR-BMA Eq. 67**	**SEM-BMA Eq. 68**	**GPR-BMA Eq. 67**	**GPR-BMA Eq. 68**
x_1	1.00	1.00	0.14	0.42	1.00	1.00
x_2	1.00	1.00	0.09	0.06	1.00	1.00
x_3	1.00	1.00	0.05	0.14	1.00	1.00
z_1	0.05	0.05	0.06	0.17	0.0	0.0
z_2	0.05	0.05	0.15	0.06	0.0	0.0
z_3	0.05	0.05	0.05	0.08	0.0	0.0
Time	77.8	88.9	74.8	68.7	933.2	713.4
Nobs	367	367	367	367	367	367
Draws	50,000	50,000	50,000	50,000	400	250

Table 1. Variable Inclusion Probabilities

Panel A. Equation 66

SAR-BMA	x_1	x_2	x_3	z_1	z_2	z_3	**Prob.**
Model 1	0	0	0	0	0	0	0.57
Model 2	0	0	0	0	1	0	0.10
Model 3	1	0	0	0	0	0	0.09
Model 4	0	1	0	0	0	0	0.05
Model 5	0	0	0	1	0	0	0.03

GPR-BMA	x_1	x_2	x_3	z_1	z_2	z_3	**Prob.**
Model 1	1	1	1	0	0	0	1.00

Panel B. Equation 67

SEM-BMA	x_1	x_2	x_3	z_1	z_2	z_3	**Prob.**
Model 1	0	0	0	0	0	0	0.35
Model 2	1	0	0	0	0	0	0.24
Model 3	0	0	0	1	0	0	0.06
Model 4	0	0	1	0	0	0	0.06
Model 5	1	0	0	1	0	0	0.06

GPR- BMA	x_1	x_2	x_3	z_1	z_2	z_3	**Prob.**
Model 1	1	1	1	0	0	0	1.00

Table 2. Selected Model Probabilities

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6
Number of Parcels	487	501	493	496	483	493
Avg. Distance (D)	0.33	0.33	0.33	0.78	0.78	0.78
Avg. Bedrooms (B)	3.0	1.5	3.0	3.0	1.5	1.5
School District (S)	1	1	0	1	1	0
Avg. Exp. Price [$E(P)$]	126.0	98.9	75.2	98.6	76.7	126.4

Table 3. Sample Summary Statistics for Simulation

Average SAR-BMA Posterior Probabilities as a Function of Sample Size

D	0.16	0.20	0.13	0.15	0.10	0.07	0.09	0.12
B	0.18	0.13	0.13	0.12	0.09	0.10	0.10	0.08
S	0.21	0.16	0.12	0.14	0.13	0.14	0.13	0.13
z_1	0.13	0.16	0.18	0.13	0.15	0.11	0.12	0.13
z_2	0.23	0.16	0.12	0.10	0.15	0.10	0.08	0.13
z_3	0.21	0.16	0.25	0.15	0.11	0.12	0.10	0.08
Time (s.)	124	123	122	122	122	122	123	123
Sample Size	60	90	120	150	180	210	240	270

Panel A. SAR-BMA Posterior Probabilities. Averaged across 10 simulation runs per sample size with 50,000 draws.

Average SEM-BMA Posterior Probabilities as a Function of Sample Size

D	0.17	0.21	0.12	0.18	0.10	0.07	0.09	0.11
B	0.21	0.13	0.11	0.13	0.10	0.19	0.15	0.10
S	0.20	0.14	0.19	0.23	0.15	0.24	0.11	0.26
z_1	0.12	0.16	0.15	0.12	0.16	0.10	0.12	0.13
z_2	0.21	0.18	0.13	0.11	0.16	0.10	0.10	0.14
z_3	0.21	0.16	0.19	0.14	0.12	0.17	0.09	0.09
Time (s.)	71	70	69	69	70	70	70	70
Sample Size	60	90	120	150	180	210	240	270

Panel B. SEM-BMA Posterior Probabilities. Averaged across 10 simulation runs per sample size with 50,000 draws.

Table 4a. Variable Selection Results

D
B
S
z_1
z_2
z_3

Average GPR-BMA Posterior Probabilities as a Function of Sample Size								
D	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
B	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
z_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
z_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
z_3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Time (s.)	142	268	302	396	411	517	693	749
Sample Size	60	90	120	150	180	210	240	270

Panel C. GPR-BMA Posterior Probabilities. Averaged across 10 simulation runs per sample size. Program terminated after convergence was achieved.

Average GWR p-values as a Function of Sample Size								
D	0.24	0.19	0.21	0.29	0.45	0.40	0.35	0.36
B	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
z_1	0.43	0.50	0.50	0.59	0.60	0.46	0.40	0.43
z_2	0.39	0.53	0.58	0.40	0.32	0.60	0.25	0.46
z_3	0.41	0.66	0.64	0.35	0.45	0.39	0.52	0.40
Time (s.)	2	2	4	5	6	7	8	10
Sample Size	60	90	120	150	180	210	240	270

Panel D. GWR p-values. Averaged across 10 simulation runs per sample size.

Table 4b. Variable Selection Results

	GPR-BMA RMSE				GWR RMSE			
Sample Size	***P***	***ME – D***	***ME – B***	**Time (s.)**	***P***	***ME – D***	***ME – B***	**Time (s.)**
60	2.14	13.52	4.41	6.91	6.83	36.85	14.57	1.91
90	1.83	11.83	3.65	14.19	5.80	31.92	16.47	2.48
120	1.57	9.69	3.20	15.58	4.58	30.63	14.87	2.99
150	1.40	8.17	3.14	23.21	3.78	27.40	13.30	3.49
180	1.32	7.58	2.91	22.94	3.36	27.26	12.92	4.01
210	1.28	7.12	2.89	28.26	3.92	28.58	13.88	5.07

Table 5. RMSE Comparisons

	GPR-BMA		GWR		GPR-BMA		GWR	
Sample Size	***ME-D***	***ME-B***	***ME-D***	***ME-B***	***ME-D***	***ME-B***	***ME-D***	***ME-B***
60	0.992	0.924	0.924	0.876	0.992	0.924	0.876	0.739
90	0.999	0.903	0.942	0.881	0.999	0.903	0.91	0.748
120	1	0.89	0.95	0.915	1	0.882	0.917	0.775
150	1	0.903	0.97	0.88	1	0.898	0.938	0.77
180	1	0.876	0.964	0.913	1	0.858	0.939	0.791
210	1	0.878	0.973	0.884	1	0.865	0.937	0.775
240	1	0.889	0.971	0.921	1	0.872	0.947	0.805
270	1	0.909	0.979	0.92	1	0.897	0.954	0.817

Table 6. Left Panel: Fraction of Significant Marginal Effects, Right Panel: Fraction of Significant Marginal Effects with Correct Sign

Variable	GPR-BMA V.I.P.	OLS-BMA V.I.P.	GPR-BMA A.M.E.	OLS-BMA A.M.E.
Latitude	0.339	0.070	0.000	0.000
Longitude	0.274	0.043	0.000	0.000
ln(GDP) in 1960	0.997	1.000	-0.011	-0.016
Life Expectancy	0.963	0.931	0.001	0.001
Equipment Investment	0.952	0.918	0.144	0.159
Non-equipment Investment	0.947	0.429	0.046	0.025
Fraction Confucian	0.917	0.988	0.051	0.056
Sub-Saharan Africa	0.755	0.744	-0.008	-0.012
Age	0.693	0.086	0.000	0.000
Fraction Hindu	0.687	0.122	-0.029	-0.003
Degree of Capitalism	0.641	0.472	0.001	0.001
Number of Years Open	0.634	0.495	0.006	0.007
Rule of Law	0.632	0.503	0.006	0.007
Latin America	0.557	0.205	-0.004	-0.002
Fraction Muslim	0.539	0.625	0.004	0.009
Fraction Buddhist	0.535	0.210	0.009	0.003
Fraction Protestants	0.508	0.466	-0.006	-0.006
Size of Labor Force	0.453	0.067	0.000	0.000
Political Rights	0.438	0.095	0.000	0.000
Black Market Premium	0.423	0.181	-0.001	-0.001
% of Pop. Speaking English	0.408	0.066	-0.002	0.000
Ratio Workers to Population	0.402	0.040	-0.002	0.000
Primary Exports	0.400	0.095	-0.003	-0.001
Fraction of GDP in Mining	0.387	0.466	0.006	0.020
Primary School Enrollment	0.366	0.206	0.000	0.004
Higher Education Enrollment	0.362	0.039	-0.018	-0.001
Fraction Catholic	0.350	0.134	-0.001	0.000
Civil Liberties	0.342	0.119	0.000	0.000
Ethnolinguistic Fractionalization	0.309	0.051	0.002	0.000
Spanish Colony	0.293	0.056	0.001	0.000
Population Growth	0.277	0.039	-0.017	0.005
War	0.256	0.078	-0.001	0.000
% of Pop. Speaking Foreign Language	0.220	0.065	0.000	0.000
French Colony	0.213	0.044	0.001	0.000
St. Dev. of Black Market Premium	0.203	0.047	0.000	0.000
Exchange Rate Distortions	0.195	0.076	0.000	0.000
British Colony	0.188	0.035	0.000	0.000
Fraction Jewish	0.186	0.037	0.001	0.000
Public Education Share	0.174	0.031	0.010	0.001
Area	0.169	0.028	0.000	0.000
Outward Orientation	0.163	0.039	0.000	0.000
Revolutions and Coups	0.129	0.028	0.000	0.000

Table 7. Variable Inclusion Probabilities and Average Marginal Effects

Country	E_Inv	NE_Inv	Country	E_Inv	NE_Inv
Malawi	0.211	0.092	*Algeria*	0.145	0.016
Cameroon	0.202	0.089	*Brazil*	0.144	0.053
Kenya	0.201	0.060	*Chile*	0.143	0.056
Tanzania	0.199	0.082	*Panama*	0.141	0.040
Nigeria	0.199	0.076	*Mexico*	0.138	0.045
Madagascar	0.198	0.104	*Costa Rica*	0.136	0.041
Ethiopia	0.196	0.110	*Argentina*	0.136	0.066
Uganda	0.195	0.099	*Taiwan*	0.135	0.042
Zimbabwe	0.191	0.069	*Portugal*	0.133	0.038
Congo	0.190	0.054	*Uruguay*	0.132	0.055
Zaire	0.189	0.098	*Venezuela*	0.129	0.039
Ghana	0.188	0.086	*Spain*	0.128	0.033
Senegal	0.187	0.084	*Cyprus*	0.124	0.006
Zambia	0.178	0.013	*India*	0.123	0.035
Philippines	0.174	0.081	*Greece*	0.120	0.004
Pakistan	0.174	0.069	*United Kingdom*	0.119	0.038
Haiti	0.169	0.097	*Hong Kong*	0.117	0.043
Morocco	0.164	0.073	*South Korea*	0.117	0.038
Thailand	0.164	0.061	*Ireland*	0.116	0.011
Bolivia	0.161	0.063	*Italy*	0.113	0.013
Tunisia	0.160	0.061	*Denmark*	0.110	0.019
Botswana	0.160	0.045	*Australia*	0.108	0.010
Turkey	0.159	0.056	*Belgium*	0.107	0.008
Honduras	0.159	0.054	*Sweden*	0.107	0.002
Sri Lanka	0.158	0.062	*Austria*	0.105	0.039
El Salvador	0.155	0.077	*Germany*	0.102	0.000
Malaysia	0.155	0.025	*Israel*	0.102	0.019
Paraguay	0.154	0.078	*Canada*	0.102	0.022
Peru	0.154	0.070	*Switzerland*	0.101	0.001
Jordan	0.153	0.045	*Netherlands*	0.101	0.007
Guatemala	0.153	0.052	*France*	0.100	0.007
Dominican Republic	0.153	0.061	*United States*	0.098	0.023
Nicaragua	0.152	0.047	*Norway*	0.097	0.000
Colombia	0.148	0.056	*Finland*	0.091	-0.015
Ecuador	0.146	0.028	*Japan*	0.075	0.001
Jamaica	0.146	0.046	*Singapore*	0.064	0.023

Table 8. Marginal Effect of Equipment and Non-Equipment Investment by Country

Appendix: Calculation of Acceptance Ratios

To calculate acceptance ratios in (37), there are two cases to consider.

1. Birth Proposals

First if $1 \le q \le k-1$ then a birth proposal is feasible, so that each transition $\delta^q \to \delta^{q+1} \in \Delta_{q+1}(\delta^q)$ is possible. Here the *acceptance ratio*, r, in (36) is given by

$$\text{(A.1)} \quad r(\delta^{q+1}, \delta^q) = \frac{p(\delta^{q+1} \mid \tilde{y}, \theta, \tilde{X})}{p(\delta^q \mid \tilde{y}, \theta, \tilde{X})} \cdot \frac{p_r(\delta^q \mid \delta^{q+1})}{p_r(\delta^{q+1} \mid \delta^q)}$$

To evaluate this expression, we note from (23) that the first ratio on the right hand side can be written as,

$$\text{(A.2)} \quad \frac{p(\delta^{q+1} \mid \tilde{y}, \theta, \tilde{X})}{p(\delta^q \mid \tilde{y}, \theta, \tilde{X})} = \frac{p(\delta^{q+1}, \tilde{y} \mid \theta, \tilde{X})}{p(\delta^q, \tilde{y} \mid \theta, \tilde{X})} = \frac{p(\tilde{y} \mid \delta^{q+1}, \theta, \tilde{X})}{p(\tilde{y} \mid \delta^q, \theta, \tilde{X})} \cdot \frac{p(\delta^{q+1} \mid \theta, \tilde{X})}{p(\delta^q \mid \theta, \tilde{X})}$$

$$= \frac{p(\tilde{y} \mid \delta^{q+1}, \theta, \tilde{X})}{p(\tilde{y} \mid \delta^q, \theta, \tilde{X})} \cdot \frac{p(\delta^{q+1})}{p(\delta^q)}$$

where the last uses the assumption that the priors of δ and θ are independent. Hence by (18) together with (20),

$$\text{(A.3)} \quad \frac{p(\delta^q)}{p(\delta^{q+1})} = \frac{p(\delta^q, q)}{p(\delta^{q+1}, q+1)} = \frac{p(\delta^{q+1} \mid q+1)p(q+1)}{p(\delta^q \mid q)p(q)} = \frac{|\Delta_{q+1}|^{-1}\, p(q+1)}{|\Delta_q|^{-1}\, p(q)} = \frac{|\Delta_q|\, p(q+1)}{|\Delta_{q+1}|\, p(q)}$$

where $p(q)$ and $p(q+1)$ are given by (19). But since

$$\text{(A.4)} \quad \frac{|\Delta_q|}{|\Delta_{q+1}|} = \frac{k!/[(q)!(k-q)!]}{k!/[(q+1)!(k-q-1)!]} = \frac{(q+1)!(k-q-1)!}{(q)!(k-q)!} = \frac{q+1}{k-q}$$

it follows from (A.2) through (A.4) that

$$\text{(A.5)} \quad \frac{p(\delta^{q+1} \mid \tilde{y}, \theta, \tilde{X})}{p(\delta^q \mid \tilde{y}, \theta, \tilde{X})} = \frac{p(\tilde{y} \mid \delta^{q+1}, \theta, \tilde{X})}{p(\tilde{y} \mid \delta^q, \theta, \tilde{X})} \cdot \frac{p(q+1)}{p(q)} \cdot \frac{q+1}{k-q}$$

Turning next to the ratio of proposal probabilities in (A.1), it follows from (34) and (35) that

$$\text{(A.6)} \quad \frac{p_r(\delta^q \mid \delta^{q+1})}{p_r(\delta^{q+1} \mid \delta^q)} = \frac{\pi(d \mid \delta^{q+1})\, p_r(\delta^q \mid d, \delta^{q+1})}{\pi(b \mid \delta^q)\, p_r(\delta^{q+1} \mid b, \delta^q)}$$

$$= \frac{\pi(d \mid \delta^{q+1})(q+1)^{-1}}{\pi(b \mid \delta^{q})(k-q)^{-1}} = \frac{\pi(d \mid \delta^{q+1})(k-q)}{\pi(b \mid \delta^{q})(q+1)}$$

Finally, substituting (A.5) and (A.6) into (A.1), we may conclude that

$$\text{(A.7)} \quad r(\delta^{q+1}, \delta^{q}) = \frac{p(\tilde{y} \mid \delta^{q+1}, \theta, \tilde{X})}{p(\tilde{y} \mid \delta^{q}, \theta, \tilde{X})} \cdot \frac{p(q+1)}{p(q)} \cdot \frac{\pi(d \mid \delta^{q+1})}{p(b \mid \delta^{q})}$$

Here the last term is seen from (30) and (31) to be identically one unless $q = k-1$, in which case $p(d \mid q+1) = 1$, and this term is equal to 2.

2. Death Proposals

Next, if $2 \le q \le k$ then a death proposal is feasible, so that each transition $\delta^{q} \to \delta^{q-1} \in \Delta_{q-1}(\delta^{q})$ is possible. In these cases the acceptance ratio is always of the form

$$\text{(A.8)} \quad r(\delta^{q-1}, \delta^{q}) = \frac{p(\delta^{q-1} \mid \tilde{y}, \theta, \tilde{X})}{p(\delta^{q} \mid \tilde{y}, \theta, \tilde{X})} \cdot \frac{p_r(\delta^{q} \mid \delta^{q-1})}{p_r(\delta^{q-1} \mid \delta^{q})}$$

But if we now let $h = q-1$ so that $h+1 = q$, then clearly $2 \le q \le k \Rightarrow 1 \le h \le k-1$. So by rewriting (A.8) in terms of h we have

$$\text{(A.9)} \quad r(\delta^{h+1}, \delta^{h}) = \frac{p(\delta^{h} \mid \tilde{y}, \theta, \tilde{X})}{p(\delta^{h+1} \mid \tilde{y}, \theta, \tilde{X})} \cdot \frac{p_r(\delta^{h+1} \mid \delta^{h})}{p_r(\delta^{h} \mid \delta^{h+1})}$$

which is seen to be exactly the *reciprocal* of (A.1) with h replacing q. So all arguments in (A.2) through (A.7) hold exactly for the reciprocals of these expressions. In particular, it now follows from (A.7) that

$$\text{(A.10)} \quad r(\delta^{h+1}, \delta^{h}) = \frac{p(\tilde{y} \mid \delta^{h}, \theta, \tilde{X})}{p(\tilde{y} \mid \delta^{h+1}, \theta, \tilde{X})} \cdot \frac{p(h)}{p(h+1)} \cdot \frac{\pi(b \mid \delta^{h})}{p(d \mid \delta^{h+1})}$$

Hence, substituting back $q-1 = h$ we may conclude that,

$$\text{(A.11)} \quad r(\delta^{q-1}, \delta^{q}) = \frac{p(\tilde{y} \mid \delta^{q-1}, \theta, \tilde{X})}{p(\tilde{y} \mid \delta^{q}, \theta, \tilde{X})} \cdot \frac{p(q-1)}{p(q)} \cdot \frac{\pi(b \mid \delta^{q-1})}{p(d \mid \delta^{q})}$$

In this case it again follows from (30) and (31) the last term is identically one unless $q = 2$, in which case $p(b \mid q-1) = 1$ and the last term is again equal to 2.