Local Marginal Analysis of Spatial Data: A Gaussian Process Regression Approach with Bayesian Model and Kernel Averaging*

Jacob Dearmon
Meinders School of Business
Oklahoma City University

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

ABSTRACT

Statistical methods of spatial analysis are often successful at either prediction or explanation, but not necessarily both. In a recent paper, Dearmon and Smith (2015) showed that by combining Gaussian Process Regression (GPR) with Bayesian Model Averaging (BMA), a modeling framework could be developed in which both needs are addressed. In particular, the smoothness properties of GPR together with the robustness of BMA allow local spatial analyses of individual variable effects that yield remarkably stable results. However, this GPR-BMA approach is not without its limitations. In particular, the standard (isotropic) covariance kernel of GPR treats all explanatory variables in a symmetric way that limits the analysis of their individual effects. Here we extend this approach by introducing a mixture of kernels (both isotropic and anisotropic) which allow different length scales for each variable. To do so in a computationally efficient manner, we also explore a number of Bayes-factor approximations that avoid the need for costly reversible-jump Monte Carlo methods.

To demonstrate the effectiveness of this *Variable Length Scale* (VLS) model in terms of both predictions and local marginal analyses, we employ selected simulations to compare VLS with Geographically Weighted Regression (GWR), which is currently the most popular method for such spatial modeling. In addition, we employ the classical Boston Housing data to compare VLS not only with GWR, but also with other well-known spatial regression models that have been applied to this same data. Our main results are to show that VLS not only compares favorably with spatial regression at the aggregate level, but is also far more accurate than GWR at the local level.

* The authors are grateful to Kelley Pace, James LeSage, Chandra Bhat and an anonymous referee for their constructive comments on an earlier version of this paper.

1. Introduction

Gaussian Process Regression (GPR) with Bayesian Model Averaging (BMA) is a powerful analytical method for modeling spatial data in contexts where little is known about either functional forms or relevant variables. In a previous paper, Dearmon and Smith (2015) [DS], compared the GPR-BMA method to a range of alternative approaches using both actual and simulated spatial data sets. This method was shown to outperform other approaches with respect to the identification of relevant predictor variables. In addition, the differentiability of GPR-BMA predictors was shown to allow local marginal estimates of individual variable effects to be calculated explicitly. While other methods can be used to study such effects, most notably Geographically Weighted Regression (GWR), it is shown below that the smooth nature of GPR-BMA predictors yields more stable estimates of these effects. Thus, the main objective of the current paper is to extend model averaging to include kernel averaging and to demonstrate its usefulness for local marginal analysis in terms of selected simulations and empirical applications. Here we focus on the well-known Boston Housing data, and show that GPR-BMA allows this data to be analyzed at a new level of spatial detail.

To do so, we start by observing that the standard isotropic version of GPR-BMA applied in [DS] is somewhat limited in terms of local marginal analysis. In particular, this model treats all variables symmetrically in terms of their influence on covariance. Thus a secondary objective of this paper is to relax this isotropy assumption in a manner that allows individual marginal influences of variables to be identified more directly (when warranted by the data). The simplest anisotropic extension of this model is to introduce separate directional length-scale effects for each variable (Section 5.1 in Rasmussen and Williams, 2006 [RW]). But there are several well-known difficulties with this extension. From a computational viewpoint, Monte Carlo estimation with individual length scales requires costly reversible-jump methods. Moreover, for cases involving many candidate variables, the simpler isotropic model with its common length scale for all variables is not only more efficient in terms of model selection, but often produces superior results. These observations suggest that the standard Bayesian method for resolving model uncertainty, namely model averaging, be extended to allow for kernel uncertainty as well. To do so, we broaden the definition of candidate models to include isotropic versus anisotropic specifications of covariance kernels. To distinguish this extended version from (isotropic) GPR-BMA, we designate the present model as the *Variable Length Scale* (VLS) model. By applying VLS to both simulated and empirical data, this extended version of Bayesian model averaging is shown to yield more robust results than either specification by itself.

A third objective of this paper is to increase the efficiency of BMA simulations in the VLS model by exploring methods for approximating Bayes factors. Such approximations not only avoid the need for costly reversible-jump methods in anisotropic cases, but more generally, require only single posterior estimates for each model considered. The best known method utilizes the Bayes Information Criterion (BIC) based on maximum-likelihood estimates of parameters. A recent scaled-prior version of this BIC approximation (SPB), proposed by Bollen et al., (2012), reweights this approximation to allow a fuller range of complex models to be considered. Our final approach involves a more direct application of the Laplace method underlying all these approximations, and utilizes maximum aposteriori (MAP) estimates of posterior mode values of parameters rather than maximum likelihood estimates. These three

approaches are compared in terms of selected simulations. Our preliminary findings here suggest that while BIC is the most efficient procedure for large sample sizes, SPB and MAP appear to be more effective in identifying true models.

To develop these results, we begin in the next section with a brief overview of Gaussian Process Regression and Bayesian Model Averaging. This is followed in Section 3 with a development of the Variable Length Scale model. In Section 4, this model is compared with Geographically Weighted Regression in terms of two selected simulation models that exhibit isotropic and anisotropic structures, respectively. Finally, this Variable Length Scale model is applied to the classic Boston Housing data in Section 5.

2. Gaussian Processes

We start with a spatial process characterized by some *response variable*, y_l , at each spatial location, l , together with a set of possible *explanatory variables*, $x_l = (x_{l1}, x_{l2}, ..., x_{lk})$, [which are implicitly taken to include spatial identifiers of location, l , such as latitude and longitude]. Our fundamental assumption is that stochastic variations in y_l over space are governed by a *zero-mean stationary Gaussian process*, which in essence implies that the joint realization of responses, $y = (y_l : l = 1,..,n)$, at any finite set of *n* locations with associated explanatory variables, $X = (x_l : l = 1,..,n)$ is *multinormally* distributed as

$$(2.1) \qquad y \sim N[0_n, c(X, X)]$$

The underlying covariance matrix,

$$(2.2) \qquad c(X, X) = \begin{bmatrix} c(x_1, x_1) & \cdots & c(x_1, x_n) \\ \vdots & \ddots & \vdots \\ c(x_n, x_1) & \cdots & c(x_n, x_n) \end{bmatrix}$$

is assumed to be generated by a *kernel function*, $c(x_l, x_h) = \mathrm{cov}(y_l, y_h)$, depending only on the attribute profiles of response variates. Spatial stationarity implies a constant mean over space, which for convenience is typically set equal to zero.[1] In essence, such processes are thus driven almost entirely by their covariance structure. In the standard *isotropic* version of this model, the covariance kernel is given by the *squared exponential function*,

$$(2.3) \qquad c_{iso}(x_l, x_h) = v \exp\left[-\tfrac{1}{2\tau^2} \| x_l - x_h \|^2\right] = v \exp\left[-\tfrac{1}{2\tau^2} \sum_{i=1}^{k} (x_{li} - x_{hi})^2\right] ,$$

which essentially hypothesizes that the covariance between responses decreases as the (Euclidean) distance between their *x*-attribute profiles increases. In particular, this implies that such covariances are *spherically symmetric* in all *x*-variables. As for the parameters of this

[1] More generally the process can be viewed as deviations about some pre-specified mean, as discussed for example in Seeger (2004).

model, note first that if $x_l = x_h$ then the exponential term collapses to unity, so that parameter v is seen to be the common variance of all responses, i.e., $\text{var}(y_l) = c(x_l, x_l) = v$. The parameter, τ, is of central importance for this kernel function, and will be discussed further below.

For our present purposes, the appropriate *anisotropic* extension of this model is given by

$$(2.4) \quad c_{aniso}(x_l, x_h) = v \exp\left[-\sum_{i=1}^{k} \frac{1}{2\tau_i^2}(x_{li} - x_{hi})^2\right] ,$$

where v has the same interpretation, and where each positive weight, $\tau_i > 0$, is designated as the *length scale* for variable x_i. In these terms, the isotropic model above is characterized by a common length scale, τ, for all explanatory variables. While length scales are often interpreted as "fluctuation rates" (where shorter length scales imply more rapid variation [RW,p.4]), the important point to stress is that variables with larger length scales must necessarily have less influence on covariance. Here it should also be stressed that such interpretations are only meaningful when variables are *standardized* to eliminate differences in measurement units. This standardization assumption will thus be implicit in all analysis to follow.

2.1 Gaussian Process Regression (GPR)

Within this Gaussian process framework, the central task has traditionally been to use a given set of data observations, $\tilde{y} = (\tilde{y}_1,..,\tilde{y}_n)$, $\tilde{X} = (\tilde{x}_1,..,\tilde{x}_n)$, to predict unobserved responses, y_l, at locations l with attributes, x_l. To do so, one additional distinction is typically made between observed and unobserved realizations of the process, namely that the observed data are also subject to observation errors. In particular, it is assumed that at each datum location, l,

$$(2.5) \quad \tilde{y}_l = y_l + \varepsilon_l \ , \ \ \varepsilon_l \sim N(0, \sigma^2) \quad , \quad l = 1,..,n$$

where y_l is the process response at *l* and where the *observation error*, ε_l, is assumed to be independent of y_l and all other observations. Thus the joint distribution of observations, $\tilde{y}$, is taken to be

$$(2.6) \quad \tilde{y} \sim N[0_n, c(\tilde{X}, \tilde{X}) + \sigma^2 I_n]$$

In these terms, the fundamental relation between observed and unobserved responses for prediction purposes is then given (as in expression (2.21) of [RW]) by

$$(2.7) \quad \begin{pmatrix} y_l \\ \tilde{y} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0_n \end{pmatrix}, \begin{pmatrix} c(x_l, x_l) & c(x_l, \tilde{X}) \\ c(\tilde{X}, x_l) & c(\tilde{X}, \tilde{X}) + \sigma^2 I_n \end{pmatrix}\right]$$

As is well known, the conditional distribution of y_l given $\tilde{y}$ must then be of the form,

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

where

(2.9) $E(y_l \mid x_l, \tilde{y}, \tilde{X}) = c(x_l, \tilde{X})[c(\tilde{X}, \tilde{X}) + \sigma^2 I_n]^{-1} \tilde{y}$

and

(2.10) $\text{var}(y_l \mid x_l, \tilde{y}, \tilde{X}) = c(x_l, x_l) - c(x_l, \tilde{X})[c(\tilde{X}, \tilde{X}) + \sigma^2 I_n]^{-1} c(\tilde{X}, x_l)$

In particular, the conditional expectation in (2.9) provides the desired mean prediction of y_l, which forms the corner stone of *Gaussian Process Regression* (GPR). More generally, the same derivation applies to vectors of unobserved responses, $y = (y_l : l = 1,..,m)$. For our later purposes it is also important to note from (2.9) that measurement-error variance, σ^2, acts formally as a "smoothing" parameter for mean predictions. In particular, if σ^2 is too close to zero, then (2.9) tends to over-fit the sample data.[2] Similarly, very large values σ^2 tend to dominate the covariance structure in (2.6), effectively over-smoothing all mean predictions in (2.9). This is of particular importance when one considers the local marginal effects of such predictions, to which we now turn.

2.2 Local Marginal Effects Analysis

To develop the local marginal effects of individual variables based on (2.9), observe first that if we employ the following simplifying notation,

(2.11) $K(\tilde{X}, \tilde{X}) = c(\tilde{X}, \tilde{X}) + \sigma^2 I_n$

and expand (2.9) as follows,

(2.12) $E(y_l \mid x_l, \tilde{y}, \tilde{X}) = c(x_l, \tilde{X}) K(\tilde{X}, \tilde{X})^{-1} \tilde{y} = [c(x_l, \tilde{x}_h) : h = 1,..,n]\, K(\tilde{X}, \tilde{X})^{-1} \tilde{y}$

then the *marginal effect* of each explanatory variable, x_i, at location l is given by the partial derivative,

(2.13) $ME_{il} = \frac{\partial}{\partial x_{li}} E(y_l \mid x_l, \tilde{y}, \tilde{X}) = [\frac{\partial}{\partial x_{li}} c(x_l, \tilde{x}_h) : h = 1,..,n]\, K(\tilde{X}, \tilde{X})^{-1} \tilde{y}$

For the *anisotropic* kernel, c_{aniso}, in (2.4) we see in particular that

[2] As is well known, GPR becomes an *exact* interpolator at data points when $\sigma^2 = 0$. This can be seen by evaluating (2.9) at all data points $(\tilde{y}, \tilde{X})$ and observing that $E(y \mid \tilde{X}, \tilde{y}, \tilde{X}) = c(\tilde{X}, \tilde{X})[c(\tilde{X}, \tilde{X}) + 0]^{-1} \tilde{y} = \tilde{y}$.

$$(2.15)\quad \frac{\partial}{\partial x_{li}} c_{aniso}(x_l, \tilde{x}_h) = v \frac{\partial}{\partial x_{li}} \exp\left[\sum_{j=1}^{k} \frac{1}{2\tau_j^2}(x_{lj} - \tilde{x}_{hj})^2\right]$$

$$= v \exp\left[-\sum_{j=1}^{k} \frac{1}{2\tau_j^2}(x_{lj} - \tilde{x}_{hj})^2\right] \frac{1}{\tau_i^2}(\tilde{x}_{hi} - x_{li})$$

$$= \frac{1}{\tau_i^2}(\tilde{x}_{hi} - x_{li}) c_{aniso}(x_l, \tilde{x}_h)$$

and similarly, for the isotropic kernel, c_{iso}, in (2.3), that

$$(2.16)\quad \frac{\partial}{\partial x_{li}} c_{iso}(x_l, \tilde{x}_h) = \frac{1}{\tau^2}(\tilde{x}_{hi} - x_{li}) c_{iso}(x_l, \tilde{x}_h)$$

These expressions will thus allow marginal effects of all variables to be calculated in closed form. Here it should be emphasized that since all variables are assumed to be standardized, the specific magnitudes of such effects are of less importance for our purposes than their signs,[3] as discussed further in Section 4.1 below.

3. Bayesian Model Averaging and Variable Selection

While the above framework implicitly treats all *k* explanatory variables, $x_l = (x_{l1}, .., x_{lk})$, as being relevant for predicting y_l, it should be clear that a fundamental part of any such analysis is to determine which variables are "most relevant". The approach adopted in [DS] (following Chen and Wang, 2010) was to treat each subset of these variables as a potential model, and to allow the observed data to reveal which of these models is most relevant. More precisely, this amounts to treating the above Gaussian process as a Bayesian prior model of responses, and then deriving the posterior probabilities of each potential model given these observations. Here we extend this approach by allowing for uncertainty about the covariance kernel of each model as well as the variables included. To formalize these ideas, we now characterize each *model*, M, as an ordered pair

$$(3.1)\quad M = (c_M, V_M)$$

where c_M is a possible *covariance kernel type* for *M*, and where $V_M \subseteq \{1, .., k\}$ denotes set of *explanatory variables* in model M. While many covariance kernel types are possible for Gaussian processes (as elaborated for example in Chapter 4 of [RW]), we focus only on the *isotropic* and *anisotropic* kernels in (2.3) and (2.4) above, so that $c_M \in \{iso, aniso\}$. If the number of explanatory variables, $s_M = |V_M|$, is designated as the *size* of model *M*, the *kernel parameters*, θ_M, of this model are seen from (2.3) and (2.4) to be specified as follows:

[3] It is often argued that standardizing allows the *relative* sizes of such effects to be more comparable (in a manner similar to the beta coefficients of standardized regression). But such interpretations have been criticized on many grounds (as summarized for example in Gelman and Pardoe, 2007), so that we choose to focus only on *directions* of change.

$$(3.2)\quad \theta_M = \begin{cases} (\nu, \tau, \sigma^2) & , c_M = iso \\ (\nu, \tau_1, .., \tau_{s_M}, \sigma^2) & , c_M = aniso \end{cases}$$

At this point it should be noted that since c_{iso} appears to be simply the special case of c_{aniso} in which all length scale parameters are the same, one may ask why the above distinction is useful. Here the key observation to be made is that even when $\tau_1 = \cdots = \tau_{s_M}$ in θ_{aniso}, the dimension of this parameter space is still $2 + s_M$ whereas the dimension of θ_{iso} is always 3, regardless of model size, s_M. As will be seen below, this implies that in cases where the data suggests that all length scales are roughly the same, c_{iso} will provide a more powerful (and computationally efficient) model than c_{aniso}. This extended notion of models forms the central element of our present *Variable Length Scale* (VLS) model.

The task remaining is to develop the full Bayesian structure of this VLS model in a manner paralleling [DS]. Here we start with the joint probability, $p(\tilde{y}, M, \theta_M)$, of any observed data vector, $\tilde{y}$, together with a parameterized model (M, θ_M). Note also that the relevant explanatory-variable data, $\tilde{X}_M$, consist precisely of those columns of $\tilde{X}$ that correspond to variables in V_M. For notational simplicity, we shall take $\tilde{X}_M$ to be implicit in every model specification. With these conventions, we now consider the probability decomposition,[4]

$$(3.3)\quad p(\tilde{y}, M, \theta_M) = p(\tilde{y} \mid M, \theta_M)\, p(\theta_M \mid M)\, p(M)$$

The conditional likelihood, $p(\tilde{y} \mid M, \theta_M)$, on the right hand side is precisely the Gaussian-process prior developed above. The conditional prior on parameters, $p(\theta_M \mid M)$, and the model prior, $p(M)$, both remain to be specified. Following standard conventions, we treat all candidate models as equally likely apriori, so that relevant distinctions between models can be attributed entirely to the observed data. Our discussion of parameter priors is deferred to the next section.

3.1 Bayes Factor Approximations

In contrast to [DS] we no longer focus on the conditional distribution, $p(\theta_M \mid \tilde{y}, M)$, obtainable from (3.3). Rather we now integrate out θ_M and focus directly on the joint distribution,

$$(3.4)\quad p(\tilde{y}, M) = \int_{\theta_M} p(\tilde{y}, M, \theta_M) d\theta_M = p(M) \int_{\theta_M} p(\tilde{y} \mid M, \theta_M)\, p(\theta_M \mid M) d\theta_M$$

which in turn implies that

[4] Following standard practice, we take *p* to represent both mass functions (such as for *M*) and density functions (such as for $\tilde{y}$).

(3.5) $$p(\tilde{y} \mid M) = \frac{p(\tilde{y}, M)}{p(M)} = \int_{\theta_M} p(\tilde{y} \mid M, \theta_M)\, p(\theta_M \mid M)\, d\theta_M$$

By evaluating these intergrals, the desired relative posterior model probabilities for any models M_1 and M_2, based on data, $\tilde{y}$, can then be obtained [under our hypothesis that $p(M_1) = p(M_2)$] from the simple identity

(3.6) $$\frac{p(M_1 \mid \tilde{y})}{p(M_2 \mid \tilde{y})} = \frac{p(M_1, \tilde{y})}{p(M_2, \tilde{y})} = \frac{p(\tilde{y} \mid M_1)}{p(\tilde{y} \mid M_2)} = \frac{\int_{\theta_{M_1}} p(\tilde{y} \mid M_1, \theta_{M_1})\, p(\theta_{M_1} \mid M_1)\, d\theta_{M_1}}{\int_{\theta_{M_2}} p(\tilde{y} \mid M_2, \theta_{M_2})\, p(\theta_{M_2} \mid M_2)\, d\theta_{M_2}}$$

These relative posterior model probabilities are thus determined entirely by the ratios, $p(\tilde{y} \mid M_1) / p(\tilde{y} \mid M_2)$, designated as the *Bayes factors* for model pairs, (M_1, M_2). However, exact evaluation of the integrals in these Bayes factors is seldom possible. Here there are several approaches. By employing appropriate conjugate priors, one can in some cases obtain tractable integral forms, as for example in Fernandez, Ley and Steel, (2001). But the most common approach is to apply Laplace's method to obtain approximations of these integrals, as developed for example in Raftery (1995) and more recently in Bollen, Ray, Zavisca, and Harden (2012) [BRZH]. In particular, a range of such approximations are summarized in [BRZH]. Here we compare three of these approximations in our simulation analyses below.

The most general approximation of (3.5), which is essentially a direct result of Laplace's method, is given in natural log terms as follows. If the number of distinct parameters in each model M is denoted by q_M (so that $q_M = 3$ for $c_M = c_{iso}$ and $q_M = s_M + 2$ for $c_M = c_{aniso}$), then,[5]

(3.7) $$\ln p(\tilde{y} \mid M) \approx \ln p(\tilde{y} \mid M, \hat{\theta}_M) + \ln p(\hat{\theta}_M \mid M) - \tfrac{q_M}{2} \ln(2\pi) - \tfrac{1}{2} \ln \det\left[-H(\hat{\theta}_M)\right]$$

where $\hat{\theta}_M$ is an estimate of the *posterior parameter mode* obtained by maximizing the first two terms on the right hand side [given data $(\tilde{y}, \tilde{X})$], i.e.,

(3.8) $$\hat{\theta}_M = \arg\max_{\theta_M} [\ln p(\tilde{y} \mid M, \theta_M) + \ln p(\theta_M \mid M)]$$

and where $\det[-H(\hat{\theta}_M)]$ is the determinant of the (negative definite) Hessian matrix

(3.9) $$H(\theta_M) = \frac{\partial^2}{\partial \theta_M \partial \theta_M'} [\ln p(\tilde{y} \mid M, \theta_M) + \ln p(\theta_M \mid M)]$$

evaluated at the maximum, $\hat{\theta}_M$. In terms of (3.8), it is natural to refer to approximation (3.7) as the *maximum-a-posteriori* (MAP) approximation.

[5] While degrees of approximation, "$\approx$", are here left unspecified, this general integral approximation is the sharpest, and is of order $O(n^{-1})$.

Note finally that this MAP approximation requires the specification of a prior distribution, $p(\theta_M \mid M)$, for θ_M. Here we start by observing that all parameters in the present framework are positive. So in the context of Gaussian processes it is most natural to assume that the logs of all parameters have normal priors as well. As in Chen and Wang (2010), we assume that parameters are independent apriori with very "diffuse" marginals, where in particular it is assumed for $c_M = c_{aniso}$ that

$$(3.10) \quad p(\theta_M \mid M) = p(v)\,p(\sigma^2)\prod_{i=1}^{s_M} p(\tau_i)$$

with log normal marginal distributions derivable from[6]

$$(3.11) \qquad \begin{aligned} \ln v &\sim N(-3,9), \\ \ln \tau_i &\sim N(-3,9), \qquad i = 1,..,s_M \\ \ln \sigma^2 &\sim N(-3,9), \end{aligned}$$

The assumptions are the same for $c_M = c_{iso}$ with τ replacing τ_i in (3.11).

There is, however, a second approximation which involves no explicit priors at all, namely the well known *Bayes information criterion* (BIC),[7]

$$(3.12) \quad \ln p(\tilde{y} \mid M) \approx \ln p(\tilde{y} \mid M, \hat{\theta}_M) - \tfrac{q_M}{2}\ln(n)$$

where the MAP estimator, $\hat{\theta}_M$, is now replaced by the *maximum likelihood estimator*,

$$(3.13) \quad \hat{\theta}_M = \arg\max_{\theta_M} \ln p(\tilde{y} \mid M, \theta_M)$$

Roughly speaking, this approximation is obtained by keeping only the largest terms in the Taylor series expansion underlying Laplace's integral approximation.

However, the BIC approximation can also be obtained by assuming a specific type of multinormal prior distribution for parameters, known as the "unit information" prior. Since this approach to BIC forms the basis for our third and final approximation, we start by noting [as in (3.11)] that multinormal priors are only reasonable for the *logs* of our positive parameters. Hence, to develop this approach, it is necessary reparameterize our model in terms of log parameter values,

---

[6] An alternative specification of priors for the kernel parameters in (2.3) and (2.4) is the gamma distribution [as discussed, for example, in Neal (1997)]. However, the log normal specifications in (3.11) are not only more convenient for our present purposes, they are also qualitatively similar to a gamma (in the present case, with shape and scale parameters of roughly 0.17 and 24, respectively).

[7] This form is taken from expression (9) in [BRZH]. As they point out, the usual expression for BIC multiplies this by -2.

$$(3.14)\quad \ln\theta_M = \begin{cases} (\ln v, \ln\tau, \ln\sigma^2) & , c_M = iso \\ (\ln v, \ln\tau_1, .., \ln\tau_{s_M}, \ln\sigma^2) & , c_M = aniso \end{cases}$$

which amounts simply to replacing each parameter value, θ_i, in (2.3) and (2.4) with the equivalent representation, $e^{\ln\theta_i}$. In these terms, the *unit information prior* for $\ln\theta_M$ is given by a multinormal distribution with any choice prior mean, $\ln\theta_o$, and with covariance matrix,

$$(3.15)\quad \text{cov}(\ln\theta_M \mid M) = \left[\tfrac{1}{n} I_o(\widehat{\ln\theta}_M)\right]^{-1}$$

based on *observed Fisher Information* evaluated at $\widehat{\ln\theta}_M$, i.e., by

$$(3.16)\quad I_o(\widehat{\ln\theta}_M) = \frac{\partial^2}{\partial \ln\theta_M \, \partial \ln\theta_M'}[-\ln p(\tilde{y} \mid M, \ln\theta_M)]_{\ln\theta_M = \widehat{\ln\theta}_M}$$

where $\widehat{\ln\theta}_M$ is obtained by replacing θ_M with $\ln\theta_M$ in the maximization problem, (3.13). Here it is important to note that this reparametrization has no effect on the resulting value for BIC in (3.12). In particular, it follows from the invariance of maximum likelihood estimators that $p(\tilde{y} \mid M, \widehat{\ln\theta}_M) = p(\tilde{y} \mid M, \hat{\theta}_M)$, and moreover that

$$(3.17)\quad \widehat{\ln\theta}_M = \ln\hat{\theta}_M$$

For our later purposes, it is also important to note that (3.16) can be calculated directly from the log-likelihood expression, $\ln p(\tilde{y} \mid M, \theta_M)$, in (3.13) without the need for explicit log transformations. In particular, it can be shown that if

$$(3.18)\quad I_o(\hat{\theta}_M) = \frac{\partial^2}{\partial \theta_M \, \partial \theta_M'}[-\ln p(\tilde{y} \mid M, \theta_M)]_{\theta_M = \hat{\theta}_M}$$

denotes the corresponding observed Fisher information matrix for the original log-likelihood, $\ln p(\tilde{y} \mid M, \theta_M)$, then each cell, $I_o(\widehat{\ln\theta}_M)_{ij}$, of the matrix in (3.16), can be calculated entirely in terms of $\theta_M = (\theta_{Mi} : 1 = 1, .., s_M)$ as follows,

$$(3.19)\quad I_o(\widehat{\ln\theta}_M)_{ij} = \hat{\theta}_{Mi}\,\hat{\theta}_{Mj}\, I_o(\hat{\theta}_M)_{ij}$$

With these preliminaries, our final approximation is motivated in [BRZH] by the observation that in many parametric settings (such as regression) the second "penalty" term in BIC tends to heavily favor models with fewer parameters (usually involving fewer explanatory variables).[8]

[8] While this appears to be less true in the present nonparametric setting, the variation of BIC proposed by [BRZH] continues to perform better than BIC in our simulation studies below.

These authors proposed a rescaling of this prior covariance which places more weight on the likelihoods of individual models (which are higher for more complex models). In particular, if we now let

$$(3.20)\quad g(\widehat{\ln\theta}_M) = (\widehat{\ln\theta}_M - \ln\theta_o)' I_o(\widehat{\ln\theta}_M)(\widehat{\ln\theta}_M - \ln\theta_o) \quad ,$$

then this rescaling yield the following *scaled-prior BIC* (SPB) approximation of $\ln p(\tilde{y} \mid M)$:

$$(3.21)\quad \ln p(\tilde{y} \mid M) = \begin{cases} \ln p(\tilde{y} \mid M, \widehat{\ln\theta}_M) - \frac{q_M}{2}\left(1 - \ln(q_M) + \ln g(\widehat{\ln\theta}_M)\right) & , \; q_M < g(\widehat{\ln\theta}_M) \\ \ln p(\tilde{y} \mid M, \widehat{\ln\theta}_M) - \frac{1}{2}\, g(\widehat{\ln\theta}_M) & , \; q_M \geq g(\widehat{\ln\theta}_M) \end{cases}$$

To be consistent with our choice of priors in (3.11) above, we here set the prior mean vector, $\ln\theta_o$, in (3.15) equal to -3 for all analyses to follow.

3.2 Metropolis-Hastings MCMC

Given these three approximations, MAP, BIC, and SPB, one could in principle simply calculate values of $p(\tilde{y} \mid M)$ for all possible models, and take the highest of these to be the model of choice. But for even modest numbers of potential explanatory variables, the number of possible models (for even a single kernel) can be prohibitive. Moreover, while there are methods for reducing this number [such as the "Occam's window" procedure of Raftery and Madigan (1994)], our present objectives are somewhat different. Here we are primarily interested in (i) identifying those explanatory variables that are statistically most relevant for responses, *y*, and (ii) estimating the local marginal effects of these variables. With respect to objective (i) in particular, note that if a given variable were to appear in many models exhibiting high model probabilities in (3.6), then even though this variable may not be included in the most probable model, it might still be very relevant for predicting *y*. So our present approach is to simulate a stochastic search process over the model space which generates model frequencies that approximate model probabilities, and at the same time allows Bayesian averaging over those models visited. This not only provides a natural statistical measure of variable relevance, but also yields more robust estimates of both the predicted responses and local marginal effects of these explanatory variables.

This procedure was developed in [DS] for models differing only in terms of included variables, i.e., $M = V_M$. So our objective here is to extend this procedure to models including kernel uncertainty as well. As above, we again suppress explanatory variables $\tilde{X}_M$ in each model specification, and let $\hat{p}(\tilde{y} \mid M)$ denote the (exponentiated) approximations of the probabilities, $p(\tilde{y} \mid M)$, for either MAP, BIC or SPB as in expressions (3.9), (3.12) and (3.21) above. If we now denote the (finite) set of possible models by

$$(3.22)\quad \mathbb{M} = \left\{ M = (c_M, V_M) : c_M \in \{iso, aniso\}, V_M \subseteq \{1, .., k\} \right\}$$

then the desired approximation of relative model probabilities for any $M_1, M_2 \in \mathbb{M}$ is given [as in (3.8)] by

$$(3.23) \quad \frac{\hat{p}(M_1 \mid \tilde{y})}{\hat{p}(M_2 \mid \tilde{y})} = \frac{\hat{p}(\tilde{y} \mid M_1)}{\hat{p}(\tilde{y} \mid M_2)}$$

Here it should be noted (from exponentiation) that all probabilities, $\hat{p}(\tilde{y} \mid M)$, are necessarily positive, so that (3.23) is well defined. In this setting, we now employ the *Metropolis-Hastings* (M-H) method to construct a Markov Chain Monte Carlo (MCMC) process on $\mathbb{M}$ with unique steady-state distribution, $\{\hat{p}(M \mid \tilde{y}, \tilde{X}) : M \in \mathbb{M}\}$.

For any Markov process, π, on $\mathbb{M}$ with transition probabilities, $\{\pi(M_j \mid M_i) : M_i, M_j \in \mathbb{M}\}$, the standard sufficient condition for $\hat{p}(M \mid \tilde{y})$ to be the unique steady state of π is that these transition probabilities satisfy the following "detailed balance" condition,

$$(3.24) \quad \hat{p}(M_i \mid \tilde{y})\pi(M_i \mid M_j) = \hat{p}(M_j \mid \tilde{y})\pi(M_j \mid M_i)$$

for all distinct $M_i, M_j \in \mathbb{M}$. Moreover, from the positivity of model probabilities, condition (3.24) in turn requires the "symmetric positivity" condition that

$$(3.25) \quad \pi(M_i \mid M_j) > 0 \Leftrightarrow \pi(M_j \mid M_i) > 0 \quad , \; M_i, M_j \in \mathbb{M}$$

The M-H approach to satisfying this condition is to decompose these transition probabilities into a *proposal process*, *pr*, and an *acceptance process*, *a*, as follows,

$$(3.26) \quad \pi(M_j \mid M_i) = pr(M_j \mid M_i)\, a(M_j; M_i)$$

In this context, if the proposal process, pr, is chosen to satisfy symmetric positivity, i.e., if

$$(3.27) \quad pr(M_i \mid M_j) > 0 \Leftrightarrow pr(M_j \mid M_i) > 0 \quad , \; M_i, M_j \in \mathbb{M}$$

and if the acceptance process is defined in terms of pr by

$$(3.28) \quad a(M_j; M_i) = \min\left\{1, \frac{\hat{p}(M_j \mid \tilde{y})}{\hat{p}(M_i \mid \tilde{y})} \frac{pr(M_i \mid M_j)}{pr(M_j \mid M_i)}\right\}$$

then it is a simple matter to verify that the detailed balance condition (3.24) is automatically satisfied. So all that remains is to construct a proposal process satisfying (3.27).

If we employ the simplifying notation, $M_i = (c_i, V_i)$, then the desired proposal process, pr, takes the form,

(3.29) $pr(M_j \mid M_i) = pr(c_j, V_j \mid c_i, V_i)$, $M_i, M_j \in \mathbb{M}$

The key point to note here is that for any set of variables, $V_i \subset \{1,..,k\}$, both kernel types $\{iso, aniso\}$ are equally well defined in terms of (2.3) and (2.4), so that either can in principle be proposed in any situation. With this in mind, the proposal procedure used here consists of two steps that are designed to allow variable switching to occur more frequently than kernel switching. In the first step, variable switching is chosen with probability .9 and kernel switching otherwise. Since only two kernel types are used, kernel switches are completely determined. For variable switching we again adopt the "birth-death" proposal process, pr_V , detailed in [DS]. As shown there, if the set, V_i, is equivalently defined by an indicator vector, δ_i, of dimension *k* with $\delta_i(h) = 1 \Leftrightarrow h \in V_i$, then the set of feasible proposal sets consists precisely of those sets, V_j, which differ from V_i by exactly one variable, i.e., for which $\Sigma_{h=1}^{k} |\delta_j - \delta_i| = 1$. From this it follows that

(3.30) $pr(c_j, V_j \mid c_i, V_i) > 0 \Leftrightarrow \{(c_j = c_i) \& (\Sigma_{h=1}^{k} |\delta_j - \delta_i| = 1)\}$ *or* $\{(V_j = V_i) \& (c_j \neq c_i)\}$

where the first bracket corresponds to *variable switching* and the second is *kernel switching*. But since all operations on the right hand side are seen to be symmetric in indices i and j (where, for example, $|\delta_i - \delta_j| = |\delta_j - \delta_i|$), it follows that the left hand side must also be symmetric, and thus that symmetric positivity condition (3.27) must hold.

3.3 Model Probabilities and Variable-Inclusion Probabilities

The above MCMC process is thus guaranteed to generate a sequence of visited models with frequencies converging to the posterior probabilities of these models. After an initial "burn in" sequence, the subsequent sequence of models, say

(3.31) $\mathcal{M} = \{M_s = (c_s, V_s) : s = 1,.., N\}$

can thus be treated as (approximate) samples from the steady-state distribution over models. If the indicator functions $\{\delta_M : M \in \mathbb{M}\}$ are defined for sequence, $\mathcal{M}$, by

(3.32) $\delta_M(s) = 1 \Leftrightarrow M_s = M$, $s = 1,.., N$

and if we again employ the notational convention in (3.23) that $\hat{p}(M \mid \tilde{y})$ denotes the approximation of $p(M \mid \tilde{y})$ under either MAP, BIC or SPB, then we may obtain natural relative-frequency estimates of these probabilities in terms of $\mathcal{M}$ as follows,

(3.33) $\hat{p}(M \mid \tilde{y}, \mathcal{M}) = \frac{1}{N} \sum_{s=1}^{N} \delta_M(s)$

where, of course, $\hat{p}(M \mid \tilde{y}, \mathcal{M}) = 0$ unless model M appears at least once in $\mathcal{M}$. In turn, one can employ (3.33) to obtain estimates of kernel probabilities. In particular, if

(3.34) $\delta_{iso}(s) = 1 \Leftrightarrow c_s = iso \ , \ s = 1,..,N$

then the *isotropy* and *anisotropy* probabilities, $\hat{p}(iso \mid \tilde{y})$ and $\hat{p}(aniso \mid \tilde{y})$, can be estimated, respectively, by

(3.35) $\hat{p}(iso \mid \tilde{y}, \mathcal{M}) = \frac{1}{N} \sum_{s=1}^{N} \delta_{iso}(s)$

and

(3.36) $\hat{p}(aniso \mid \tilde{y}, \mathcal{M}) = 1 - \hat{p}(iso \mid \tilde{y}, \mathcal{M})$

Of more importance for our present purposes are estimates of posterior *variable-inclusion probabilities*, $\hat{p}(i \mid \tilde{y})$, for each variable, $i = 1,..,k$, which can be obtained using variable indicators,

(3.37) $\delta_i(s) = 1 \Leftrightarrow i \in V_s$

as follows,

(3.38) $\hat{p}(i \mid \tilde{y}, \mathcal{M}) = \frac{1}{N} \sum_{s=1}^{N} \delta_i(s)$

As in [DS], these inclusion probabilities provide a natural statistical measure of relevance for each variable, which (unlike p-values) is *larger* for more relevant variables.

3.4 Bayesian Model Averaging of Predictions and Marginal Effects

Finally, the model sequence, $\mathcal{M}$, can also be used to obtain more robust estimates of both predictions and marginal effects. By again employing the same "hat" notation for a representative approximation method, MAP, BIC or SPB, we now let $\hat{\theta}_s$ denote the relevant parameter estimates of θ_{M_s} for each model M_s visited under that approximation method. So in particular, $\hat{\theta}_s$ denotes the posterior mode estimates in (3.8) under MAP, and denotes the maximum likelihood estimates in (3.12) and (3.21), respectively, under BIC and SPB (where for notational simplicity we take $\hat{\theta}_s$ to represent the log value, $\widehat{\ln \theta_s}$ in SPB).

If the corresponding predictions at each location l are denoted by

(3.39) $\hat{E}(y_l \mid x_l, \tilde{y}, \tilde{X}, M_s, \hat{\theta}_s) = c_{M_s}(x_l, \tilde{X}_{M_s}) K(\tilde{X}_{M_s}, \tilde{X}_{M_s})^{-1} \tilde{y}$,

then the resulting *BMA prediction* based on model sequence, $\mathcal{M}$, is given by

(3.40) $\hat{E}(y_l \mid x_l, \tilde{y}, \tilde{X}) = \frac{1}{N}\sum_{s=1}^{N} \hat{E}(y_l \mid x_l, \tilde{y}, \tilde{X}, M_s, \hat{\theta}_s)$

Similarly, if the corresponding marginal effect of each variable, x_i, $i = 1,..,k$, at location l is defined (as in [DS]) by

(3.41) $\widehat{ME}_{il}^{\,s} = \begin{cases} \frac{\partial}{\partial x_i}\hat{E}(y_l \mid x_l, \tilde{y}, \tilde{X}, M_s, \hat{\theta}_s) & , i \in V_{M_s} \\ 0 & , i \notin V_{M_s} \end{cases}$

then the resulting *BMA marginal effect* based on model sequence, $\mathcal{M}$, is given by

(3.42) $\widehat{ME}_{il} = \frac{1}{N}\sum_{s=1}^{N} \widehat{ME}_{il}^{\,s}$

As mentioned in the introduction, these BMA marginal effects will form the main focus of our subsequent analysis. So it is of particular importance to develop some measure of "posterior credibility" for at least the directions (signs) of these effects. But the posterior distributions of both mean responses, and their partial derivatives are only obtainable in terms of the posterior distributions of the parameters, θ_M , defining these values. Moreover, unlike the two-stage MCMC procedure in [DS] which generates samples of θ_M from this posterior distribution, all parameters here have been integrated out in (3.4) and (3.5). So there is no direct sampling mechanism for approximating posterior distributions of marginal effects.

But in a manner similar to our asymptotic approximations of posterior model probabilities (using MAP, BIC, or SPB), it is possible to obtain asymptotic approximations of the posterior distribution of θ_M , and to sample marginal effects based on this distribution. Here our initial approach was to follow the general recommendations in Gelman, et al. (2013, Section 13.3) by applying a multinormal approximation to the log posterior, $\ln \theta_M$. However, experimentations with such approximations revealed that local marginal effects are particularly sensitive to measurement-error variance, σ^2 (as mentioned at the end of in Section 2.1 above). In particular, variations in such effects for any given model, *M*, tend to be magnified by only small deviations from the maximum likelihood estimate, $\hat{\sigma}_M^2$. So for purposes of the present analysis, we choose to treat $\hat{\sigma}_M^2$ as the optimal value of this smoothing parameter, and consider only variations in the *kernel parameters*, $\theta_M^c = \theta_M - \{\sigma^2\}$. In this setting, both the posterior mean and mode of $\ln \theta_M^c$ are identical under asymptotic normality, and are thus asymptotically approximated by $\widehat{\ln \theta}_M^c = \ln \hat{\theta}_M^c$. Similarly, (following Gelman et al., 2013) posterior covariances for unimodal distributions are generally well approximated by the inverse of *observed Fisher Information* (i.e., by the local curvature of the log likelihood around the mode). This yielding a posterior approximation of the form,

(3.43) $\ln \theta_M^c \underset{approx}{\sim} N\left[\ln \hat{\theta}_M^c, \mathrm{cov}(\ln \hat{\theta}_M^c)\right]$

where $\ln\hat{\theta}_M^c$ is simply the sub-vector of $\ln\hat{\theta}_M = (\ln\hat{\theta}_M^c, \ln\hat{\sigma}_M^2)$ in (3.17), and where $\text{cov}(\ln\hat{\theta}_M^c)$ is the corresponding sub-matrix of the inverse, $I_o(\ln\hat{\theta}_M)^{-1}$, of the observed Fisher Information in (3.19) [given $\hat{\sigma}_M^2$].

By drawing samples, $\ln\theta_M^c$, from these multinormal distributions and transforming back to the model parameters, $\theta_M = (\theta_M^c, \hat{\sigma}_M^2)$ with $\theta_M^c = \exp(\ln\theta_M^c)$, we can construct corresponding samples from the posterior distributions of both mean predictions (3.39) and marginal effects (3.41) for each visited model, M. For marginal effects in particular, if we now write $M \in \mathcal{M}$ whenever model M appears in sequence $\mathcal{M}$, and if we draw, say, $S_M = 1000$ samples from (3.43) and transform to parameter samples $(\theta_M^s : s = 1,..,S_M)$, then in terms of (3.41) above, we can construct corresponding marginal-effect samples

$$(3.44) \quad ME_{ilM}^s = \begin{cases} \frac{\partial}{\partial x_i} E(y_l \mid x_l, \tilde{y}, \tilde{X}, M, \theta_M^s) & , i \in V_M \\ 0 & , i \notin V_M \end{cases}$$

Similarly, we can employ (3.39) to construct prediction samples, and can in principle then construct standard credibility intervals based on these sample frequencies. But here our main interest focuses on marginal effects, where such credible intervals are somewhat more difficult to interpret directly (in view of the standardization of explanatory variables mentioned at the end of Section 2). Thus to evaluate these effects, we begin by observing that standardizations have no influence on the *signs* of such effects. Moreover, since the overall relevance of individual variables has already been gauged in terms of variable inclusion probabilities (VIPs), it is the signs (directions) of their local effects on mean responses that are of most interest. With this in mind, our main objective is to construct a measure of the "sign credibility" of estimated marginal effects.

While many approaches are in principle possible here, the most sensible (to us) is simply to estimate the posterior probability of such signs. In particular, if in terms of (3.44) we now let $p_M(ME_{ilM})$ denote the posterior probability distribution of marginal effects, ME_{il} , in model M, and let S_{ilM}^+ (resp., S_{ilM}^-) denote the number of positive (resp., negative) values of samples, ME_{ilM}^s, in (3.44) then the natural posterior estimates of these signs are given by

$$(3.45) \quad \hat{p}_M(ME_{ilM} > 0) = \frac{S_{ilM}^+}{S_M} \quad , \quad \text{and}$$

$$(3.46) \quad \hat{p}_M(ME_{ilM} < 0) = \frac{S_{ilM}^-}{S_M}$$

i.e., by the fractions of sampled marginal effects that are positive (or negative). Finally, by using the estimated model probabilities in (3.33), we can then obtain the following overall (BMA) estimates of *posterior sign probabilities*:

(3.47) $\hat{p}(ME_{il} > 0) = \sum_{M \in \mathcal{M}} \hat{p}_M(ME_{ilM} > 0)\, \hat{p}(M \mid \tilde{y}, \mathcal{M})$, and

(3.48) $\hat{p}(ME_{il} < 0) = \sum_{M \in \mathcal{M}} \hat{p}_M(ME_{ilM} < 0)\, \hat{p}(M \mid \tilde{y}, \mathcal{M})$

While it is evident that this procedure for determining "sign credibility" involves a number of assumptions, our simulations below show that it does indeed provide reasonable results. In particular, the estimated signs of marginal effects for relevant variables turn out to have sign probabilities exceeding .95 in most cases where these effects are not too close to zero.

4.0 Simulation Analyses

In this section we develop two simulation models that are specifically designed to test the ability of the VLS model to detect differences between spatial isotropic and anisotropic processes. The general form of these simulated processes is as follows,

(4.1) $y_l = \phi(x_{l1}, x_{l2}) + \varepsilon_l$, $\varepsilon_l \underset{iid}{\sim} N(0, 0.25)$

where the use of only two explanatory variables in the mean-value function, ϕ , allows results to be mapped and analyzed visually as well as numerically. In addition, these mean-value functions are chosen to be continuously differentiable to allow local marginal analyses to be tested against actual partial derivatives at each location. As in [DS], the random spatial errors, ε_l are thus chosen to be sufficiently small [$\text{var}(\varepsilon_l) = 0.25$] to ensure that these functional specifications always dominate residual noise.[9]

Within this framework, our *isotropic* simulation model is chosen to have the following mean-value function,

(4.2) $\phi_{iso}(x_{l1}, x_{l2}) = 2\sin(2x_{l1}) + 2\sin(2x_{l2})$

which is clearly symmetric in x_{l1} and x_{l2} . So there should be no discernible difference between the length scales of these two variables as identified by VLS. In addition, the local marginal effects of both x_1 and x_2 are identical, and are given by

(4.3) $\frac{\partial}{\partial x_i}\phi_{iso}(x_{l1}, x_{l2}) = 4\cos(2x_{li})$, $i = 1, 2$

The *anisotropic* simulation model has the more complex form

(4.4) $\phi_{aniso}(x_{l1}, x_{l2}) = 2\sin(2x_{l1}) + \sin\left(\frac{1}{2}x_{l2}\right) + \sin(4x_{l1})\sin(x_{l2})$

[9] The spatial errors in the simulations of [DS] were also chosen to be spatially autocorrelated. But this additional refinement turned out to make no discernible difference in the results, and has now been dropped.

involving an interaction term. Note also from the first term that values of x_1 are more influential than x_2. In particular, the local marginal effects of each variable are now given by

(4.5) $\frac{\partial}{\partial x_{l1}}\phi_{aniso}(x_{l1}, x_{l2}) = 4\cos(2x_{l1}) + 4\cos(4x_{l1})\sin(x_{l2})$

(4.6) $\frac{\partial}{\partial x_{l2}}\phi_{aniso}(x_{l1}, x_{l2}) = \frac{1}{2}\cos(\frac{1}{2}x_{l2}) + \sin(4x_{l1})\cos(x_{l2})$

In our simulation analyses below, it is important to compare VLS 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.

4.1 Geographically Weighted Regression

Since this method is available in many software packages, we simply use the "off the shelf" version that is currently available in ArcMap. Moreover, since there is an extensive literature on this technique [most notably, the monograph by Fotheringham, Brunsdon and Charlton (2002)] we sketch only those aspects needed for our present comparison.[10] In terms of the observed data $(\tilde{y}, \tilde{X})$ given above, this approach directly estimates both predictions and local marginal effects at each location, l, by means of a weighted regression scheme of the form

(4.7) $\min_{\beta_l} \sum_{j=1}^{n} (\tilde{y}_j - \tilde{x}'_j \beta_l)^2 w_l(j)$

where the *spatial kernel* weights, $w_l(j)$, emphasize those locations, *j*, closest to location, l. To facilitate the present comparison with VLS, the spatial kernel is chosen to have the *squared exponential* form,

(4.8) $w_l(j) = \exp(-d_{lj}^2 / b^2)$

which is seen to parallel our present use of the squared exponential covariance kernel in (2.3) and (2.4) above (and which is also the default choice in ArcMap). However, it is important to emphasize that while "distances" in the covariance kernels of VLS involve *all* explanatory variables, distances, d_{ij}, in (4.8) involve only the spatial coordinates of locations i and j. Finally, the parameter $b > 0$ is usually designated as the *bandwidth* of the kernel, and plays a role somewhat analogous to length scales (restricted to spatial coordinates). The choice of an appropriate bandwidth is typically carried out by standard "leave-one-out" cross-validation techniques.[11]

[10] Here it should be noted that GWR appears to have been introduced independently by both Brunsdon, Fotheringham, and Charlton (1996) and McMillen (1996). So while our present approach (using ArcMap software) is based on the former approach [as summarized in their 2002 monograph above], the reader is also referred to the excellent papers by McMillen (2010, 2012) and McMillen and Redfearn (2010).

[11] See for example Brunsdon, Fotheringham, and Charlton (1996, Section 3.2).

Given this basic GWR model, the desired beta coefficients, $\beta_l = (\beta_{l1},..,\beta_{lk})'$, are estimated by ordinary least squares to obtain the closed-form solution:

(4.9) $\hat{\beta}_l = (\tilde{X}'W_l\tilde{X})^{-1}\tilde{X}'W_l\tilde{y}$

where W_l is a diagonal matrix with components, $w_l(j)$, $j = 1,..,n$. With these estimates, the *mean response prediction*, $E(y_l \mid x_l, \tilde{y}, \tilde{X})$, at location l is simply the standard regression prediction,[12]

(4.10) $\hat{E}(y_l \mid x_l, \tilde{y}, \tilde{X}) = \tilde{X}\hat{\beta}_l = \sum_{j=1}^{n} \hat{\beta}_{lj} x_{lj}$

So the desired *local marginal effects* of variable, x_i, at location l are simply the estimated beta coefficients themselves, i.e.,

(4.11) $\frac{\partial}{\partial x_i}\hat{E}(y_l \mid x_l, \tilde{y}, \tilde{X}) = \frac{\partial}{\partial x_i}\sum_{j=1}^{n} \hat{\beta}_{lj} x_{lj} = \frac{\partial}{\partial x_i}(\hat{\beta}_{li} x_{li}) = \hat{\beta}_{li}$

In this setting, the "sign credibility" of such marginal effects is now replaced by p-values for local "pseudo t-tests" of these betas.[13] In doing so, however, it should be emphasized that such tests are well known to be unstable in the presence of either small samples $(n < 150)$ or multicollinearity, where the usual convention (as for example in ArcMap) is to require that the condition number (ratio of maximum and minimum eigenvalues) of the postive semidefinite matrix $\tilde{X}'W_l\tilde{X}$ not exceed 30. But since $n = 150$ in the simulations used for comparisons below, and since almost all condition numbers turn out to be well less than 30,[14] the present simulations appear to provide a reasonably fair comparison with GWR.

4.2 Simulation Results

For all simulations conducted, the range of values chosen for the explanatory variables in (4.2) and (4.4) was $-2 \le x_i \le 2$, $i = 1,2$. A series of 12 sample sizes were chosen, ranging from $n = 40$ to $n = 150$ in increments of 10. For each sample size, a set of 20 simulations were conducted in which *n* values of x_1 and x_2 were randomly drawn from the interval [-2,2]. Values of *y* were then simulated from models (4.2) and (4.4). For testing purposes, three irrelevant variables (x_3, x_4, x_5) were also drawn randomly from [-2,2], and both VLS and GWR were then applied to each of these data sets $\{(y_i, x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}) : i = 1,..,n\}$. However, our initial investigations

[12] See Harris, Fotheringham, Crespo, Charlton (2010) for further elaboration on GWR predictions.
[13] Most tests focus on the overall (average) relevance of explanatory variables in this context, such as the nonparametric test developed by McMillen and Redfearn (2010). However, since practitioners often tend to rely local pseudo t-tests for betas [as for example in Mennis and Jordan (2005) and Matthews and Yang (2012)] which do appear in other software [such as GWR4 (https://geodacenter.asu.edu/gwr_software)], we choose this method for purposes of comparison. (See footnote 21 below for further discussion.)
[14] Only one of the 1681 local predictive regressions has a condition number above 30 (and this with value 30.92).

showed that GWR was so sensitive to the presence of irrelevant variables, that a reasonable comparison was only made possible by dropping the irrelevant variables from the GWR applications.[15] In contrast, such variables had almost no effect on VLS results. In fact the VIPs in Table 4.1 below show that for the post burn-in sample sequence, $\mathcal{M}$, in (3.31) above, these irrelevant variables were virtually eliminated by the Metropolis-Hastings procedure.

The same out-of-sample prediction set was used for each simulation, and consisted of a regular grid of points on the square $[-2,2]\times[-2,2]$, spaced at intervals of 0.1. This produced a set of $1681(=41^2)$ prediction points which (as seen in Figures 4.1) allowed the prediction surfaces to be interpolated and plotted in full detail. With these preliminary observations, we turn now to the results of these simulations.

Predictions and Variable Relevance

The prediction results for both Sim 1 (isotropy) and Sim 2 (anisotropy) can be seen most vividly by examining *y*-predictions for typical simulation runs at sample size, $n=150$, as shown in Figure 4.1 below (using the SPB approximation method).[16] Here plots of the true mean-value functions in (4.2) and (4.4) are shown in the middle panels, (b) and (e), of Figure 4.1, respectively. The corresponding prediction results for GWR are shown on the left [panels (a) and (d)], and those for VLS are shown on the right [panels (c) and (f)]. Here it is seen that both GWR and VLS are able to capture the overall surface pattern quite well, especially in the simpler isotropic case (which is symmetric in the two explanatory variables). However, it is also clear that VLS does a noticeably better job of capturing the finer details, especially for the more complex anisotropic case. This contrast will be even more dramatic in the marginal-effect results below.

Figure 4.1

But for the present, we focus on certain more detailed aspects of VLS, and in particular, on the relative performance of the three Bayes-factor approximation methods (MAP, BIC, SPB). The first question of interest relates to how well these three methods predict the *y* response in Sim 1 and Sim 2. In Figure 4.2 below, these results are summarized for each sample size in terms of Root Mean Squared Error (RMSE), where the results shown are averaged over the 20 simulations at each sample size. Here all three methods are seen to be roughly comparable, and exhibit the same degree of improvement at larger sample sizes.[17]

Figure 4.2

[15] Such problems with GWR are well known, and the ArcMap documentation recommends that preliminary OLS regressions be done in an attempt to identify and remove irrelevant variables.

[16] We use Vanhatalo et. al.'s (2013) GPstuff package to conduct GPR runs and D'Errico's (2014) adaptive robust numerical differentiation for Hessian evaluation.

[17] The only exception here is the one outlier case at sample size $n = 60$ in Sim 2, where SPB is performing noticeably worse than both MAP and BIC. Here it appears that larger numbers of simulations would be warranted.

Next we observe that VLS successfully identifies the isotropic and anisotropic nature of Sim 1 and Sim 2, respectively, but that again, both SPB and MAP do substantially better than BIC in identifying the isotropy in Sim 1.

Figure 4.3

In terms of length scales, VLS does equally well in distinguishing between the isotropic and anisotropic cases. As seen in Figure 4.4 below (using only SPB)[18], the estimated length scales of both explanatory variables, x_1 and x_2, are almost identical, as expected. In addition, the more influential nature of x_1 in Sim 2 is here reflected by a substantially shorter length scale than x_2 across all sample sizes. While length-scale results for the irrelevant variables are not shown, it suffices to say that for all irrelevant variables (x_3, x_4, x_5), length scales were much larger than those of x_2 in Figure 4.4 for all sample sizes (and all models for which that variable was included).

Figure 4.4

But much sharper (and more easily interpretable) results for variable relevance are given in terms of variable inclusion probabilities, as shown in Table 4.1 below (for three representative sample sizes). In particular these results show that for both Sim 1 and Sim 2, the relevant variables (x_1, x_2) were present in virtually all models of the post burn-in sequences, $\mathcal{M}$. However, there are considerable differences between the three Bayes-Factor approximation method (MAP, BIC, SPB), with respect to the irrelevant varaibles, (x_3, x_4, x_5). In particular, BIC exhibits by far the worst performance in this respect, and certain irrelevant variables are included in more than 30% of the models in $\mathcal{M}$. Here it is clear that in almost all cases, SPB is doing the best in this regard (especially in the more complex anisotropic case where irrelevant variable inclusion is reduced to less than 1% at larger sample sizes).

Table 4.1

Because we are only concerned with local marginal effects of the most relevant variables, we have adopted SPB as the default choice for the empirical analysis in Section 5 below. In addition, it is for this reason that the SPB approximation was used in constructing Figure 4.1 above, together with Figure 4.5 below.

[18] Here both BIC and MAP yield almost identical results to SPB.

Local Marginal Effects

Finally we turn to a consideration of the local marginal effects that form the central focus of the present paper. As a parallel to Figure 4.1 above, we begin with a graphical comparison (Figure 4.5 below) of both VLS and GWR marginal effects for Sim 2 with the true marginal effects of variables x_1 and x_2 in expressions (4.5) and (4.6), respectively.[19] Here we again focus on the same representative sample of size 150 (using SPB) in Figure 4.1 above.

Figure 4.5

Notice first that the true marginal effects in this anisotropic model are substantially different between x_1 and x_2. In particular, it can be seen from the vertically elongated contours of ϕ_{aniso} in panel (e) of Figure 4.1 that the marginal effect (ME2) of x_2 is relatively small, with more fluctuations around zero.[20]

For the most important variable, x_1, we see from panel (a) of Figure 4.5 that GWR is still able to capture the overall qualitative nature of the true marginal effect pattern (ME1) in panel (b). But it is again clear that VLS in panel (c) is able to pick up the finer details of this pattern with much greater precision. However, the most dramatic differences between these two methods are seen when comparing their performance on variable, x_2. Here it is evident from panel (d) of Figure 4.5 that GWR is simply not able to detect the more subtle (and rapid) variations of these marginal effects (ME2).[21] In marked contrast, the true pattern in panel (e) continues to be reflected in the VLS results of panel (f). So this example serves to underscore the differences in sensitivity between these two approaches. Note finally that even VLS is unable to detect the very intricate pattern variations near the upper boundary of panel (e), and like all interpolation methods, is limited by pattern "edge effects".

These results are also supported by the corresponding significance and credibility levels for estimates in each case. Turning first to GWR, rather than complicating these figures with exact contours of significance levels, we have simply plotted local *t*-values in Sim 2 against their corresponding beta values x_1 and x_2 in panels (a) and (b) of Figure 4.6 below. Here we have also included horizontal lines to indicate the sets of significant values for the standard "default" range, $|t| \geq 2$.[22] Here it is seen that betas for unstable case of x_2 are insignificant over a much

[19] Recall that the (essentially one-dimensional) marginal effects for the separable symmetric model in Sim 1 are far less interesting from a spatial viewpoint, and will be considered only in terms of relative fit (for VLS) below.

[20] The shapes of these ME2 contours can best be seen by visually verifying that the contour-tangency points on any vertical line in the ϕ_{aniso} plot [panel (e) in Figure 4] correspond precisely to the ME2 = 0 contour, which is the second lowest (coolest) contour in panel (e) of Figure 4.5.

[21] Examinations of different sample patterns for this case produce the same qualitative results for GWR, and in addition, exhibit rather extreme variations from sample to sample.

[22] Such default levels of significance are no doubt inflated, and should in principle be corrected by estimating "effective degrees of freedom" [as is typically done for tests of average beta values in GWR (Leung, 2000)]. So in the present setting such levels are best viewed as simply a benchmark for comparison with VLS.

broader range of beta values than for x_1 . Even more telling is the fact (hard to guage from the figure) that while 86.3% of the beta values for x_1 are significant, only 27.2% of the beta values for x_2 are significant. So these default levels of significance provide some indication that the results for x_2 should be interpreted with caution. However, it should also be noted that the extremely negative values of beta for x_2 are all very "significant", even though they are way below the actual minimum partial slope value of -0.5.

Figure 4.6

To compare these significance results with VLS, the sign probabilities of each marginal effect for variables x_1 and x_2 are plotted in panels (c) and (d) of Figure 4.6. Here the dots above the horizontal lines at .95 identify values that are credible at the 95% level or above. Note that (as stated at the end of Section 3.4) the tapering of these scatter plots around zero shows that less credible marginal effects tend to be those with values closer to zero. However, what is more difficult to guage from the figure is the overwhelming concentration of values at the upper end. In fact, the percentages of marginal effect values with sign probabilities above .95 are 83.5% for x_1, and 72.2% for x_2. So with respect to x_2 in particular, these results are seen to be consistent with the qualitatively better fit in panel (f) than in panel (d) of Figure 4.5.

Finally, as with prediction and variable relevance above, it remains to consider the relative marginal-effect performance of our three Bayes-factor approximation methods (MAP, BIC, SPB), as shown in Figure 4.7 below. Here again it is clear that these estimates are roughly comparable across all methods.[23] This is especially true at larger samples sizes ($n \geq 80$), where all three methods again produced much sharper results.[24]

Figure 4.7

5. Empirical Application: Boston Housing

In this final section we apply VLS to a standard spatial data set, namely the Boston housing data first studied by Harrison and Rubinfeld [HR] (1978) . These authors employed multiple regression (OLS) to identify the effect of "demand for clean air" on median housing prices for each of the 506 census tracts in Boston (at the time of the 1970 census). But the first explicit attempt to analyze this data from a spatial perspective was that of Pace and Gilley (1997), who collected centroid data for these tracts and applied a spatial errors model with weight matrix based on centroid distances. More recently, Deng (2008) has analyzed this same dataset using a

[23] For purposes of RMSE comparisons with (4.3),(4.5) and (4.6), marginal effects of (dimensionless) standardized variates have been rescaled [multiplied by their sample standard deviations] to approximate their original dimensions.

[24] Note that the effects of the single outlier case for n = 60 in Sim 2 of Figure 4.2 are reflected by the corresponding marginal results of panel (b) in Figure 4.7. But here, MAP seems to have been slightly affected as well as SPB.

spatial lag model in which the underlying weight matrix explicitly incorporates certain types of anisotropies based on this data. So one objective of our present analysis is to compare the variables found to be significant in these two spatially-oriented studies with those found to be most relevant in terms of our present variable-inclusion probabilities. However, our main objective is to explore this data set at a finer level of spatial detail by focusing on the local marginal effects of relevant explanatory variables.

To do so, we begin by summarizing the original [HR] model, with the *median value* (MV) of housing in each tract as the dependent variable, and with *nitrogen-oxide* (NOX) representing the air-quality variable of interest. The additional 13 control variables include *particulate concentrations* (PART), *average number of rooms* (RM), *proportion of structures built before 1940* (AGE), *black population proportion* (B), *lower status population proportion* (LSTAT), *crime rate* (CRIM), *proportion of area zoned with large lots* (ZN), *proportion of nonretail business areas* (INDUS), *property tax rate* (TAX), *pupil-teacher ratio* (PTRATIO), *location contiguous to the Charles River* (CHAS), *weighted distances to the employment centers* (DIS), and an index of *accessibility to radial roads* (RAD). As with many regressions, a number of these variables were transformed for purposes of analysis.[25] Subsequently, Pace and Gilley (1997) added the spatial coordinates, *latitude* (LAT) and *longitude* (LON), together with a full quadratic specification of these two variables (i.e., LAT^2, LON^2, LAT*LON) for their spatial analysis. In this light, one important additional feature of VLS is that such nonlinearities are implicitly captured by this flexible nonparametric model. So in the analysis to follow, the original variables (including LAT and LON) will be used for VLS analyses. The only variable that requires further comment is black proportion, B. If for convenience we now use BP to denote the *proportion* (*fraction*) *of tract population that is Black*, then the actual specification of explanatory variable, B, above is $B = (BP - .63)^2$. However, it was possible to recover the original variable, BP, from the 1970 census, and it is this variable that is used in our present analysis.

5.1 Global Predictions and Variable Relevance

Here we begin with global results including both mean predictions of median housing prices and identification of relevant variables. The relevant data on median housing prices for Boston census tracts is shown on the left panel in Figure 5.1 below (and is one of the first *spatial* representations of this classical data set that has appeared).[26] The right panel shows the mean predictions generated by VLS using SPB,[27] which are almost indistinguishable from the original data (at this level of choropleth approximation). So even though these are necessarily *in-sample* predictions,[28] they do provide empirical support for the predictive accuracy of this method, as seen in the simulation studies above.

---

[25] The variables MV, LSTAT, DIS, and RAD were transformed to natural logs. In addition, NOX was transformed to NOX^2 , and B was transformed as discussed in the text.

[26] The present shapefile of 1970 Boston census tracts was extracted from the digital boundary files of the National Historical Geographic Information System (NHGIS). A similar shapefile was developed by Roger Bivand, and is now available as part of the *spdep* package in R.

[27] Results for MAP are very similar here and are not shown.

[28] Note that *out-of-sample* predictions are of limited use in studies of this type, where the given 506 census tracks form contiguous spatial units. However, for studies involving geocoded point data (such as individual housing sales), it is clear that out-of-sample predictions would be far more appropriate.

Figure 5.1a

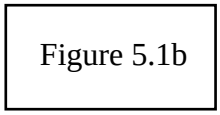


Figure 5.1b

But unlike the simulation examples above, it is more difficult to compare these prediction results with GWR. The main difficulty here relates to the nature of this Boston housing data itself, which in many cases is spatially far coarser that individual census tract units. In particular, for the key explatory variable, NOX, there are only 81 distinct values. Moreover, these are in contiguous patches (shown in Fig 5.2a below for the central Boston area) that constitute a higher level of aggregation than census tracts.[29]

Figure 5.2a

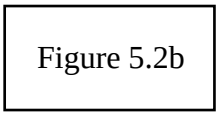
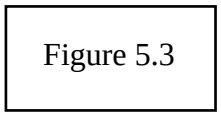


Figure 5.2b

The situation is even worse for variables like RAD, where each Town in the Boston area was assigned a unique value (as shown for the central Boston area in Fig 5.2b). The local collinearties created by such variables make it very difficult to apply GWR to this dataset.[30] So for purposes of comparison with VLS, we have employed a more robust version of GWR constructed by James LeSage [31], which employs pseudo inversion to circumvent problems of singular (or almost singular) matrices. This allows approximate predictions to be constructed as in (4.8) and compared with those of VLS. The results of this comparison are shown for the central Boston area in Figure 5.3 below (which is the area of most interest for our subsequent analysis). In particular, prediction residuals for VLS (using SPB) and GWR are shown for each tract in panels (a) and (c),[32] respectively, with true median values in panel (b).[33] This local comparison is sufficient to indicate that even with robust pseudo-inversion techniques, the predictions obtained by GWR are far less reliable. Moreover, even for those explanatory variables exhibiting distinct values for each tract, the beta estimates obtained by pseudo-inversion are generally not interpretable. So in the local analysis of marginal effects below, we will focus only on VLS.

Figure 5.3



Given these general prediction results, we turn next to the identification of relevant variables. Recall that for VLS, such identification is in terms of variable inclusion probabilities (VIP). In Table 5.1 below, we compare these VIPs in panel (b) [using both SPB and MAP] with the spatial

[29] As discussed in Bivand (2015), this NOX data is based on measurements in 122 meteorological (TASSIM) zones that were in turn "copied out" to more aggregate collections of census tracts.

[30] In ArcMap, for example, the condition-number restrictions mentioned in Section 4.1 above only allow GWR to be run with small subsets of the Boston variables.

[31] The GWR program, **gwr.m**, is part of his Matlab *Econometrics Toolbox* (Version 7, 2012).

[32] The common residual scale for VLS and GWR is in median-value units with *white* = ±2, and with increments of 2 units on either side.

[33] In this enlarged map it can also be seen that a number of census tracts are missing (including downtown Boston itself). As noted by Bivand (2015), this is in part due to missing data or to small number of housing units in these tracts.

regression estimates and significance levels for each variable obtained by both Pace-Gilley and Deng in panel (a). As mentioned in the introduction to Section 5, none of the transformations of variables (in footnote 28) were used in VLS. So the variable labels in panels (a) and (b) differ in this respect. For convenience, these variables are ordered in terms of their VIP values (for SPB) from highest to lowest. As shown by the horizontal lines in the tables, there is close agreement between variables significant at the .05 level in panel (a) and variables with VIP ≥ 0.95 in panel (b). The only exceptions here are RAD and PTRATIO (the latter for SPB only), with VIPs well below .95. Like RAD in Figure 5.2 (b), PTRATIO is also constant in each town, so that both represent categorical variables that are spatially step functions. So while VLS is clearly more robust than GWR with respect to such variables, they can in some cases conflict with the continuity assumption underlying VLS (and in particular its GPR component). Finally, to compare the directions of influence for each variable, we have also included the average values of Marginal Effects (Avg. ME) across all 506 census tracts.[34] Here there is essentially complete agreement in signs between the two panels.

Table 5.1

5.2 Local Marginal Effects

With respect to local marginal effects, we begin by observing that such effects are far more tenuous for variables at higher levels of spatial aggregation than census tracts. In particular, the NOX variable of most interest to [HR] falls into this category. So rather than attempt to reaggregate such data at the level of "NOX zones", we focus here on several variables that are well defined at the census-tract level. The two variables we have chosen to examine are Black Proportions (BP) and AGE.

Local Analysis of BP Effects

As mentioned above, this variable is of particular interest because it was originally specified as a quadratic expression, $(BP - .63)^2$. As stated by [HR], this was done in order to capture not only the general negative effect of BP on median housing values, but also its slight positive effect in areas with very high concentrations of Black population. One may ask why a quadratic term in BP was not used in their regression. The answer seems to be that (even with standard zero centering of the quadratic term) this nonlinear effect is not significant. Only by prespecifying an appropriately calibated form can such a small effect be picked up. So in the present setting, it is of interest to ask whether such slight nonlinear effect can be identified by the marginal effects of BP estimated in VLS. To do so, we have plotted in Figure 5.4 all BP values (in percent terms) against their marginal effect values, ME_BP, obtained from an application of VLS.[35]

Figure 5.4

[34] Note that the zero values for LAT and CHAS reflect the fact that neither of these variables were included in any model of sequence $\mathcal{M}$.

[35] As stated above, all VLS applications in this section use the SPB approximation of Bayes Factors.

Here we have included only those ME values that are credible at the 95% level (eliminating mostly values close to zero). While the vast majority of these credible BP marginal effects are negative, there are a couple of tracts with very high BP and positive ME_BP. These two tracts (805 and 818) are in South Boston where Blacks are most heavily concentrated, and are thus fully consistent with the hypothesis of [HR]. So this small effect does indeed seem to be picked up by VLS. But much more interesting is the cluster of positive marginal effects in the upper left-hand corner of Figure 5.4. This represents the opposite extreme where Black proportions are among the lowest – a phenomenon that seems to have been missed by [HR]. To gain further insight, we focus on one of these tracts (3542), as shown by the small circle in Figure 5.4. This tract, which is just west of Havard Square in Cambridge (as shown on the left in Figure 5.5 below) has the highest 1970 median housing prices in this data set, and has a Black proportion of less than 2%. (The red color of this tract on the right in Figure 5.5 reflects its high positive ME_BP value, while blue shades denote tracts with negative ME_BP values). As for the socio-economic profile of this tract, more than half of all adults (25 and older) were college graduates in 1970. This, together with its proximity to Harvard, suggest that most heads of households were upper-middle-income professionals. If so, then the positive marginal effect of BP found here might well reflect a very different population dynamic than in South Boston.[36] In any case, this example serves to illustrate how VLS can identify local variations in the spatial fabric that allow deeper levels of analysis.

Figure 5.5

Local Analysis of AGE Effects

An even more dramatic example is provided by the AGE variable (as measured by the proportion of housing units builts before 1940). Here again we focus only on the central area of Boston, where some of the oldest residential areas can be found. As seen in Figure 5.6b below, most of the marginal effects for AGE are negative (shades of blue), indicating that that higher proportions of older homes tend to detract from median value. However, aside from the pink area around MIT (with values so close to zero that none exceed 80% credibility),[37] there is seen to one striking exception right in the heart of Boston's South End district. This census tract (709), with 80% Black population and all houses built before 1940, appears to be something of an anomaly – with a positive marginal AGE effect in the top 3% of all census tracts. However, further investigation reveals that this area contains the single "largest urban Victorian neighborhood in the country".[38] Moreover, the South End Historical Society was created in 1966 precisely to preserve this area, as shown by the brown boundary in Figure 5.6a (later designated as the Boston Landmark District). Tract 709 is seen to lie entirely inside this area, as shown by the red boundary in Figure 5.6b. Such preservation efforts led in turn to an influx of middle income families seeking to purchase and restore these houses – which may very well have been reflected in the 1970 census. So it would appear that these VLS estimates of local marginal AGE

[36] By way of contrast, fewer than 40% of all adults in tracts 805 or 818 were even high school graduates.
[37] Aside from this MIT area, all other marginal effects in Figure 5.6b are credible at the 95% level.
[38] See for example "A Short History of Boston's South End" by Arlene Vadum (available online at *http://www.south-end-boston.com/History*).

effects again reveal meaningful spatial variations that are by no means apparent from more conventional methods of analysis.

Figure 5.6

6. Concluding Remarks

In this paper we have extended the basic GPR-BMA framework in [DS] to include alternative kernel choices for candidate models, as well as Bayes-factor approximations to expedite the process of stochastic model search in this extended framework. In particular, the inclusion of anisotropic kernels with individual length scales for candidate variables allows sharper identification of their local marginal effects. By employing selected simulations, this Variable Length Scale (VLS) model was shown to yield predicted responses and local marginal effects that are far more accurate than the currently prevailing method for doing so, namely Geographically Weighted Regression (GWR). In terms of Bayes-factor approximations (BIC,SPB,MAP) for VLS, it was found that both SPB and MAP tended generally to outperform BIC. However, this difference was noticeably reduced for larger samples sizes. This together with its relative computational simplicity, suggest that BIC may well turn out to be the better choice for larger sample sizes.

In addition, both VLS and GWR were applied to the classical Boston Housing data, where VLS was shown to yield far more reliable (in-sample) prediction. In addition, VLS was shown to yield VIP results (together with average marginal-effect estimates) that are qualitatively very comparable to results obtained by previous spatial regression analyses of this data. In terms of local marginal analyses, where spatial regression methods cannot be applied, the lumpy nature of the Boston data also precluded any meaningful estimation by GWR. So the more robust nature of VLS is made fully evident by this application. In particular, several specific examples were developed to show that VLS allows deeper levels of local analysis than is possible by these other methods.

However, a number of research directions remain to be explored. Chief among these is the pragmatic issue of scalability. Even with Bayes-factor approximations, the VLS model continues to be computationally very costly (orders of magnitude more costly than either spatial regression or GWR). One approach here is to reduce the number of prior candidate models for VLS, using methods such as the “Occam’s window” procedure of Madigan and Raftery (1994). More promising, perhaps, are the numerous data-reduction techniques currently used for reducing the costs of large matrix calculations. These range from simple random subsampling of the data to more complex methods of identifying “best representative subsamples”, such as the recent “variational Bayes” approaches of Hensman et al. (2013) and Gal et al. (2014).

Aside from computational efficiency, we are also exploring extensions of VLS that would broaden the application domain of this model. One direct extension would be to increase the range of covariance kernels considered. More generally, such extensions might include

relaxations of spatial stationarity in terms of "multiple kernels" with possible interactions similar to Lloyd et al. (2014).

REFERENCES

Bivand, R. (2015) "Revisiting the Boston data set (Harrison and Rubinfeld, 1978): a case study in the challenges of system articulation", *Working Paper, SAM 30* (ISSN: 0804-6824) Department of Economics, Norwegian School of Economics, Helleveien 30, N-5045 Bergen, Norway.

Bollen, K.A., S. Ray, J. Zavisca, and J.J. Harden (2012) "A Comparison of Bayes Factor Approximation Methods Including Two New Methods", *Sociological Methods & Research*, 41(2): 294-324.

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

Dearmon, J. and T.E. Smith (2015) "Gaussian Process Regression and Bayesian Model Averaging: An alternative approach to modeling spatial phenomena." Forthcoming in *Geographical Analysis.*

D'Errico, J. (2014). Adaptive Robust Numerical Differentiation - File Exchange - MATLAB Central. Retrieved January 16th, 2015, from http://www.mathworks.com/matlabcentral/fileexchange/13490-adaptive-robust-numerical-differentiation

Deng, M. (2008) "An anisotropic model for spatial processes." *Geographical Analysis*, 40(1): 26-51.

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

Gal, Y., M. van der Wilk, and C. Rasmussen (2014) "Distributed variational inference in sparse Gaussian process regression and latent variable models", *Advances in Neural Information Processing Systems*, *Paper 1660*, Montreal, Canada.

Gelman, A. and I. Pardoe (2007) "Average predictive comparisons for models with nonlinearity, interactions, and variance components", *Sociological Methodology* 37: 23-51.

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

Harrison, D. and D.L. Rubinfeld (1978) “Hedonic housing prices and the demand for clean air”, *Journal of Environmental Economics and Management*, 5: 81-102.

Hensman, J., N. Fusi, and N.D. Lawrence (2013) “Gaussian processes for big data”, *arXiv preprint arXiv:1309.6835*.

Leung, Y. (2000) “Statistical tests for spatial nonstationarity based on the geographically weighted regression model”, *Environment and Planning A,* 32: 9-32.

Lloyd, J.R., D. Duvenaud, R. Grosse, J.B. Tenenbaum, and Z. Ghahramani (2014)"Automatic construction and natural-language description of nonparametric regression models." *arXiv preprint arXiv:1402.4304.*

Madigan, D. and A.E. Raftery (1994) “Model selection and accounting for model uncertainty in graphical models using Occam's window”, *Journal of the American Statistical Association,* 89:1535-46.

Matthews, S.A and T. Yang (2012) “Mapping the results of local statistics: using geographically weighted regression”, *Demographic Research*, 26: 151–166

McMillen, D. (2010) “Issues in spatial data analysis”, *Journal of Regional Science*, 50: 119-141.

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.

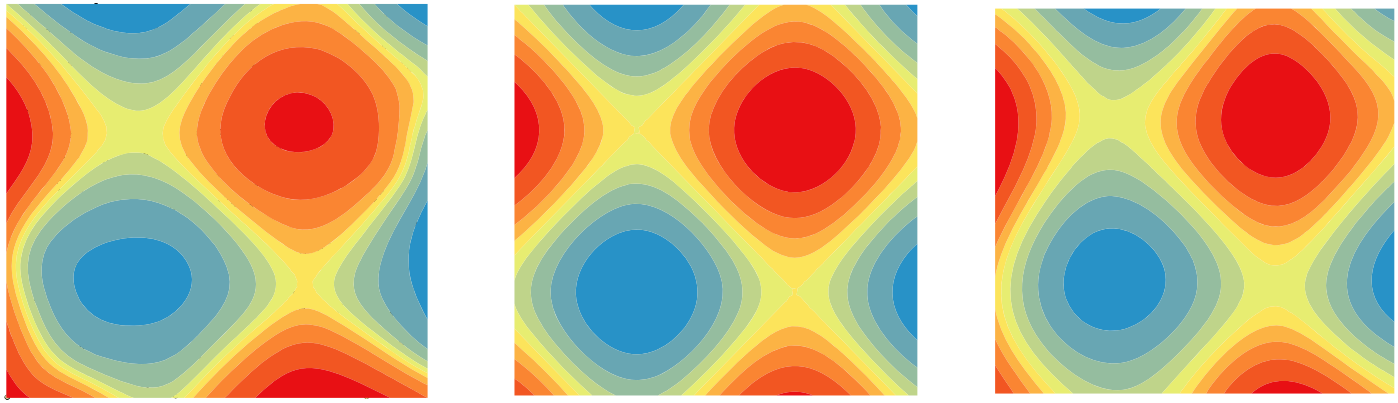
Mennis, J.L. and L.M. Jordan (2005) “The distribution of environmental equity: exploring spatial nonstationarity in multivariate models of air toxic releases”, *Annals, Association of American Geographers*, 95:249–268.

Pace, R.K. and O.W. Gilley (1997) “Using the spatial configuration of data to improve estimation”, *The Journal of Real Estate Finance and Economics*,

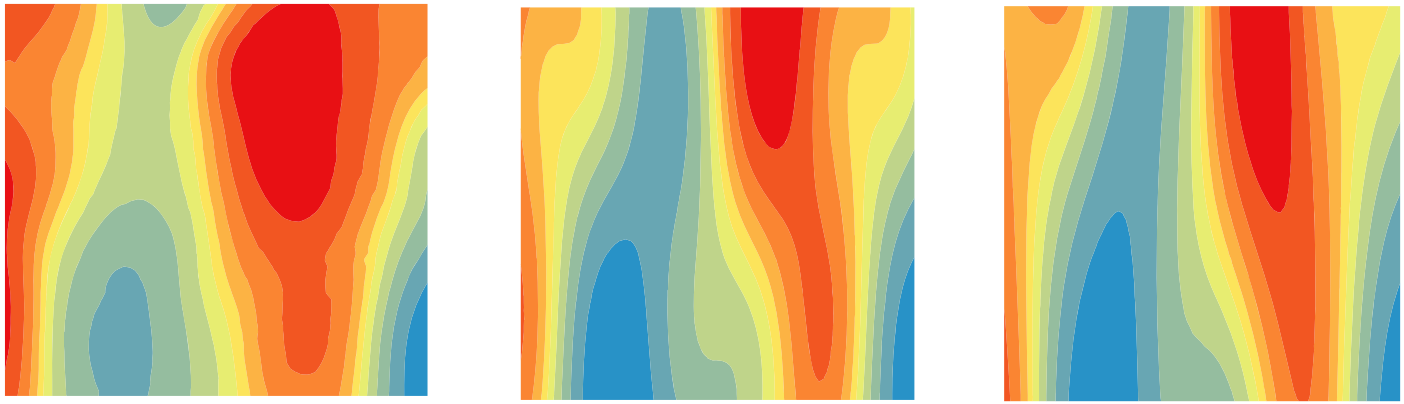
Seeger, M. (2004) “Gaussian processes for machine learning”, *International Journal of Neural Systems*, 14: 69-106.

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., & Vehtari, A. (2013). GPstuff: Bayesian modeling with Gaussian processes. *The Journal of Machine Learning Research*, *14*(1), 1175-1179.

Figures



(a) SIM1_Y_GWR (b) SIM1_Y_True (c) SIM1_Y_VLS



(d) SIM2_Y_GWR (e) SIM2_Y_True (f) SIM2_Y_VLS

Figure 4.1. Prediction Results for Simulations

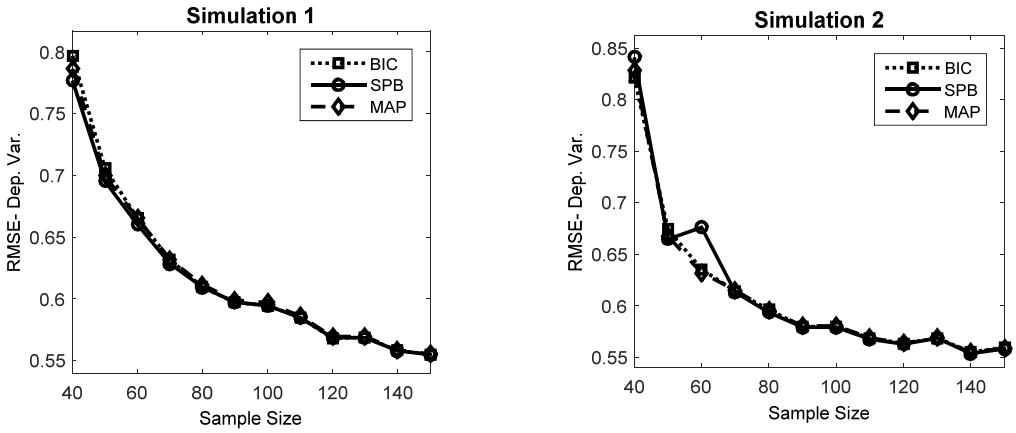


Figure 4.2. Comparative Prediction Fits

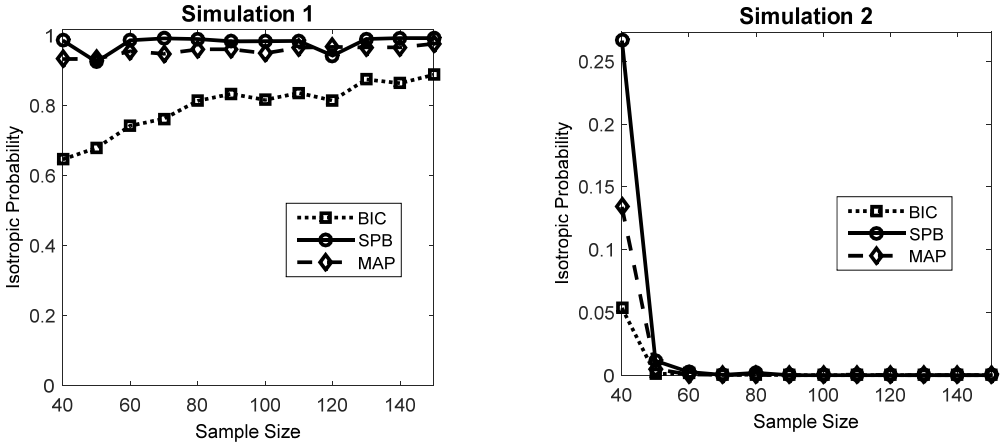


Figure 4.3. Comparative Prediction Fits

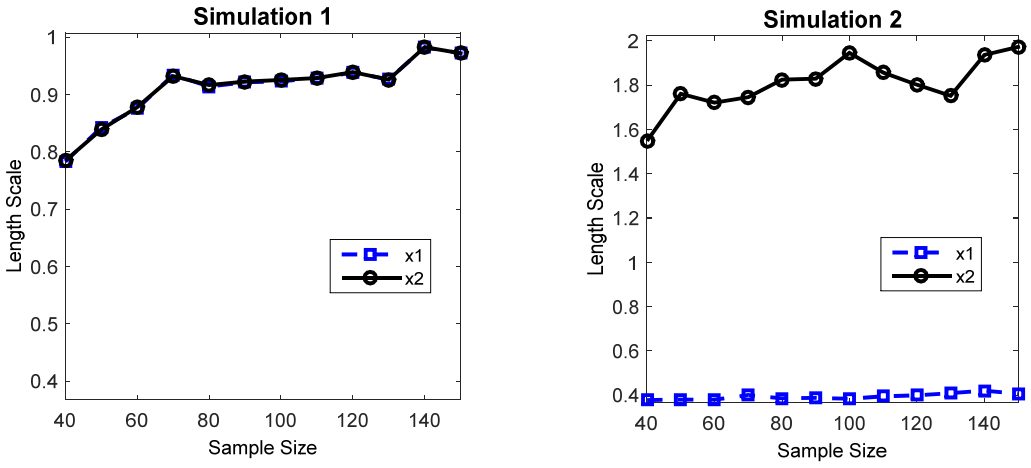
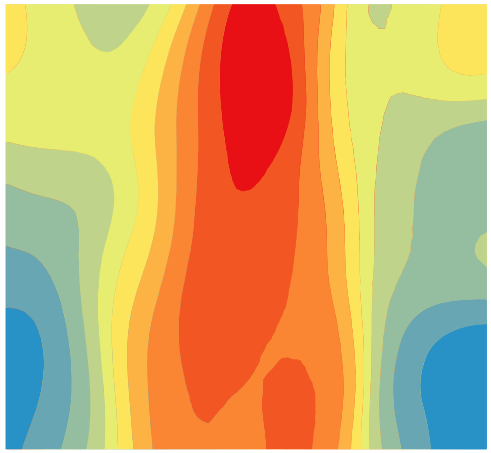
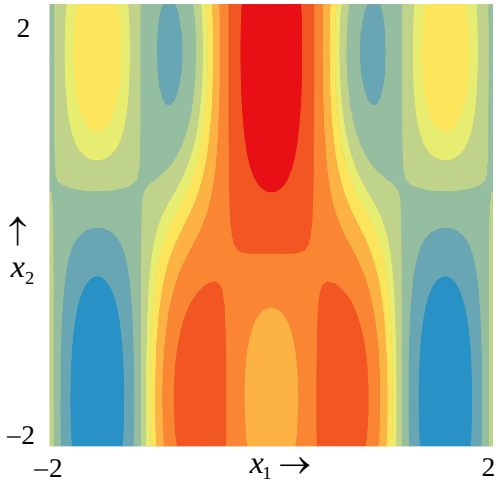


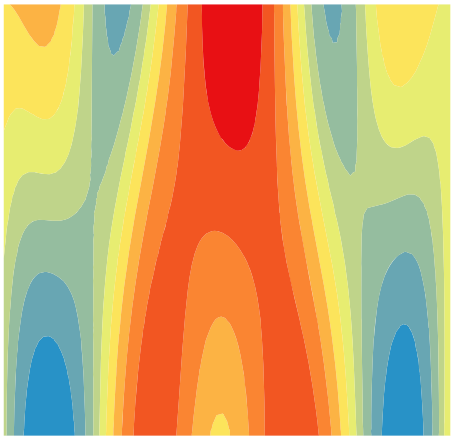
Figure 4.4. Length Scale Comparisons



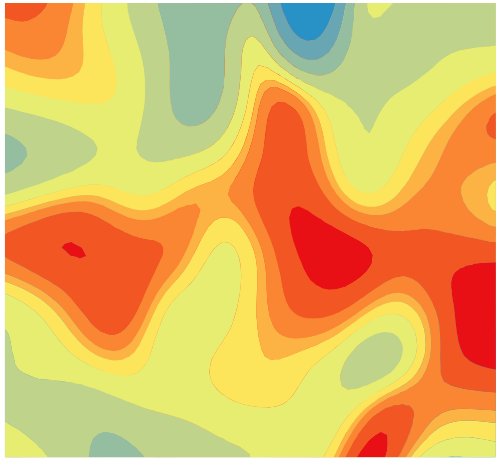
(a) SIM2 ME1 GWR



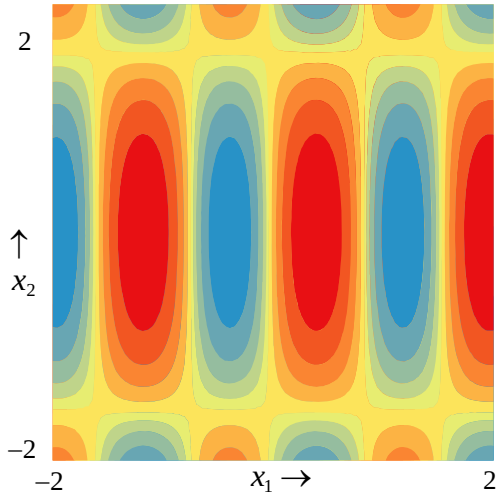
(b) SIM2 ME1 True



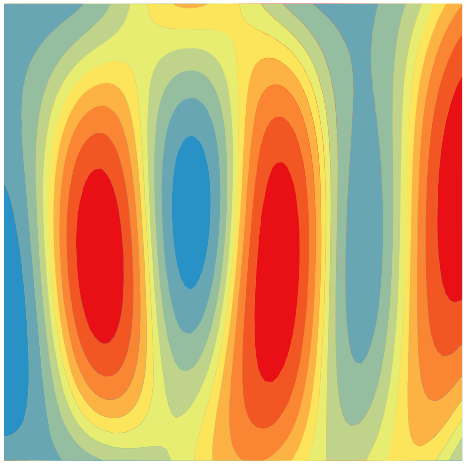
(c) SIM2 ME1 VLS



(d) SIM2 ME2 GWR

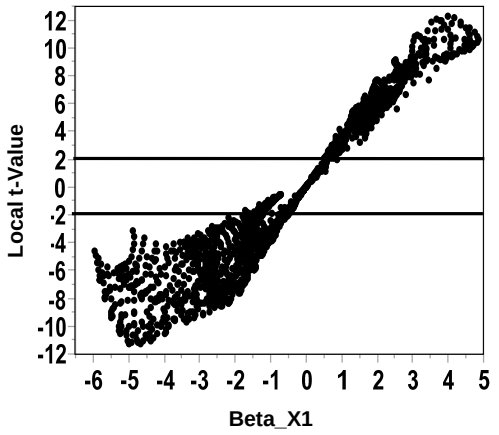


(e) SIM2 ME2 True

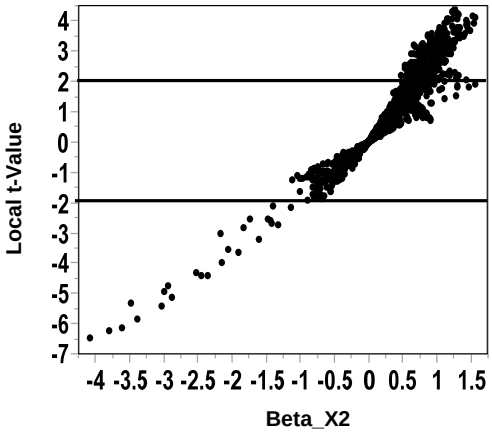


(f) SIM2 ME2 VLS

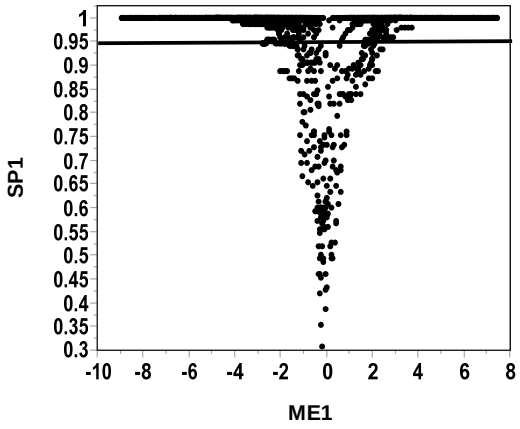
Figure 4.5. Comparison of Marginal Effects



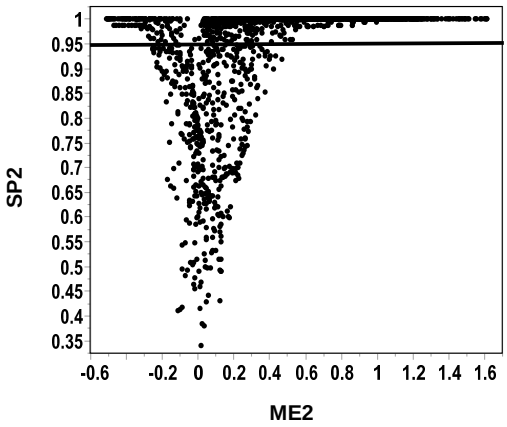
(a) Local t-Values for X1 Beta



(b) Local t-Values for X2 Beta

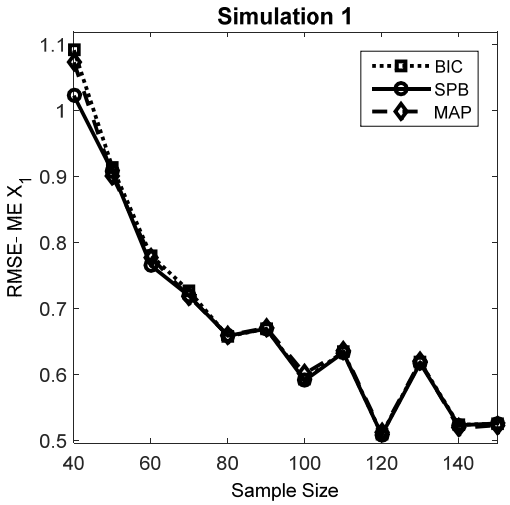


(c) Sign Probabilities for ME1

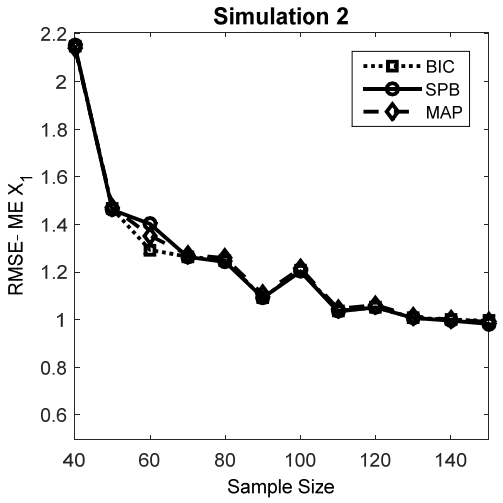


(d) Sign Probabilities for ME2

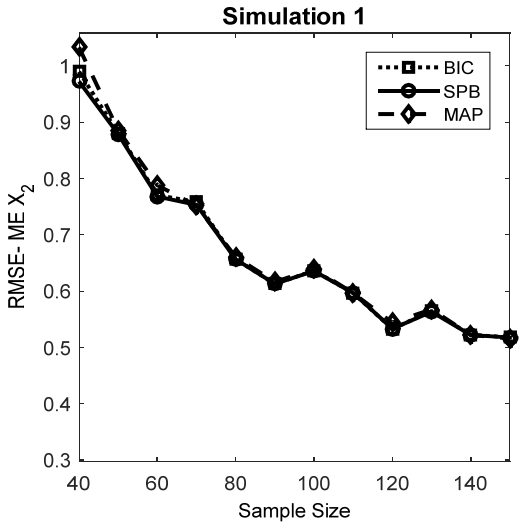
Figure 4.6. GWR Significance versus VLS Credibility



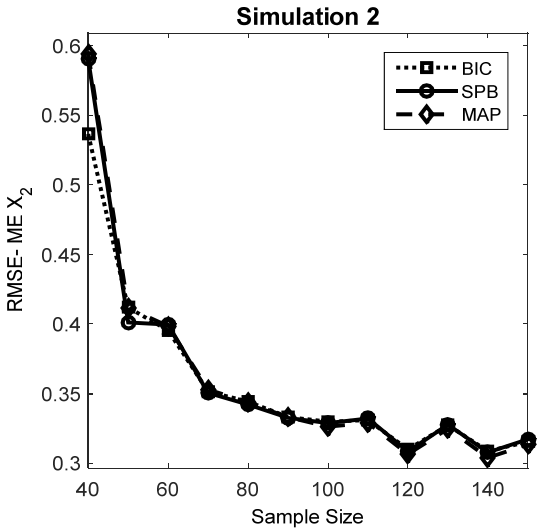
(a) ME_X1 for SIM 1



(b) ME_X1 for SIM 2

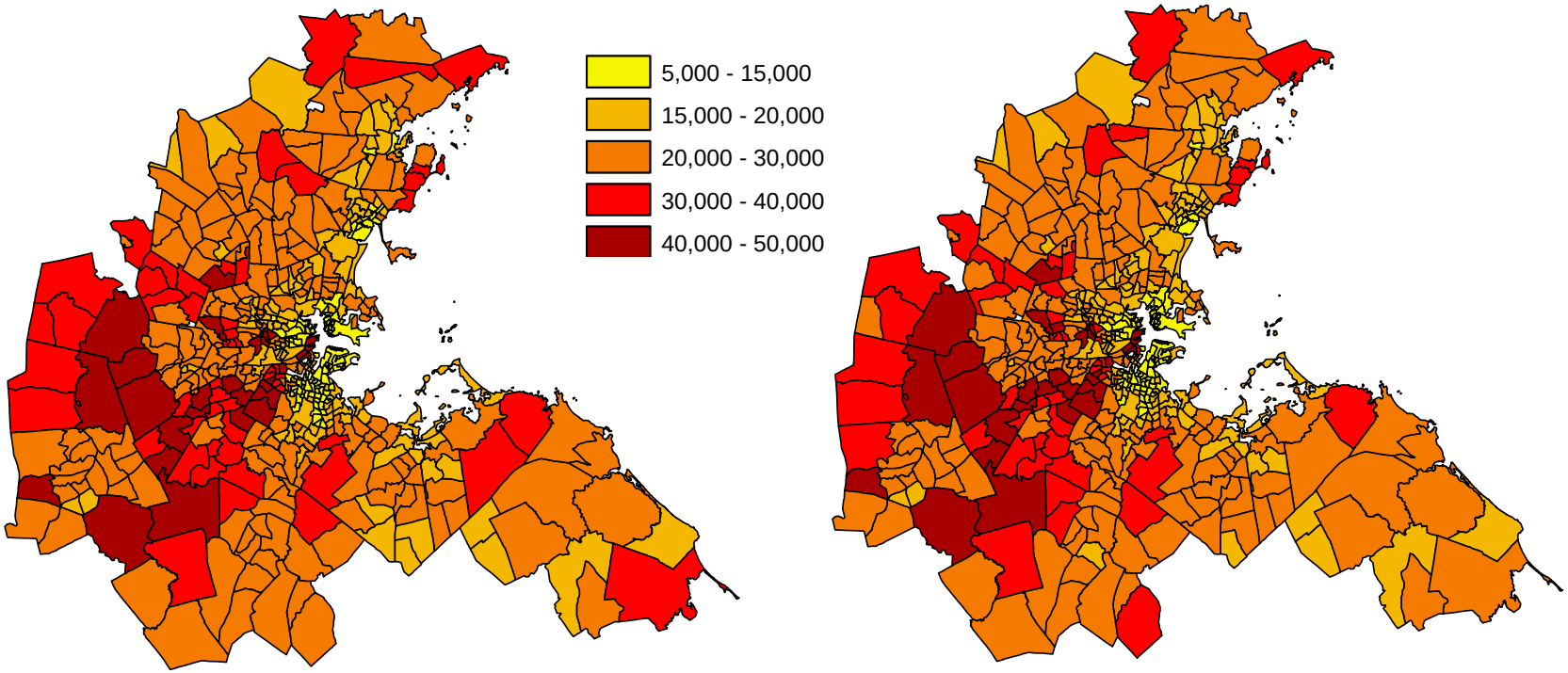


(c) ME_X2 for SIM 1



(d) ME_X2 for SIM 2

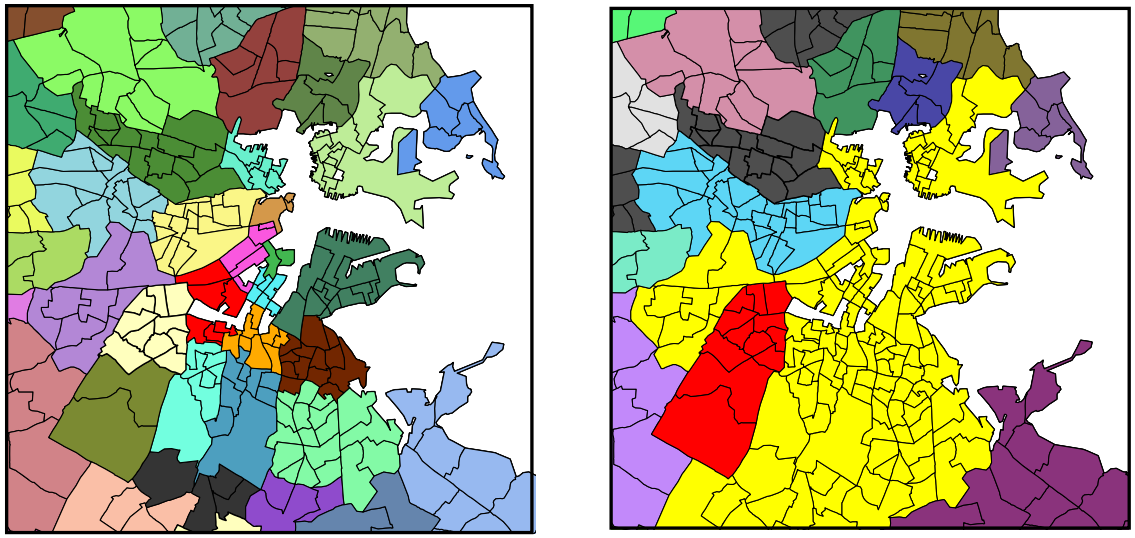
Figure 4.7. Comparisons of Local Marginal Effects



(a) Median Values 1970

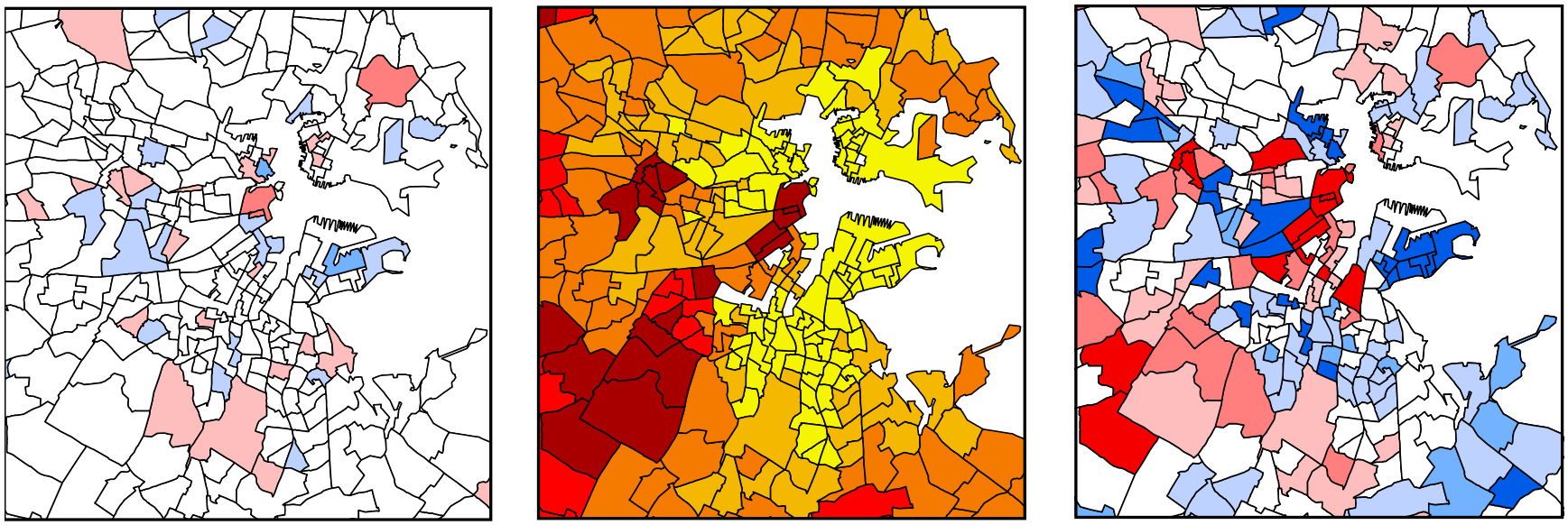
(b) Predicted Median Values

Figure 5.1. Comparison of Median Values with Predictions



(a) NOX Constancy (b) RAD Constancy

Figure 5.2. Variables with Higher Aggregation Levels



(a) VLS Residuals

(b) True Median Values

(c) GWR Residuals

Figure 5.3. Residual Comparison for Central Boston

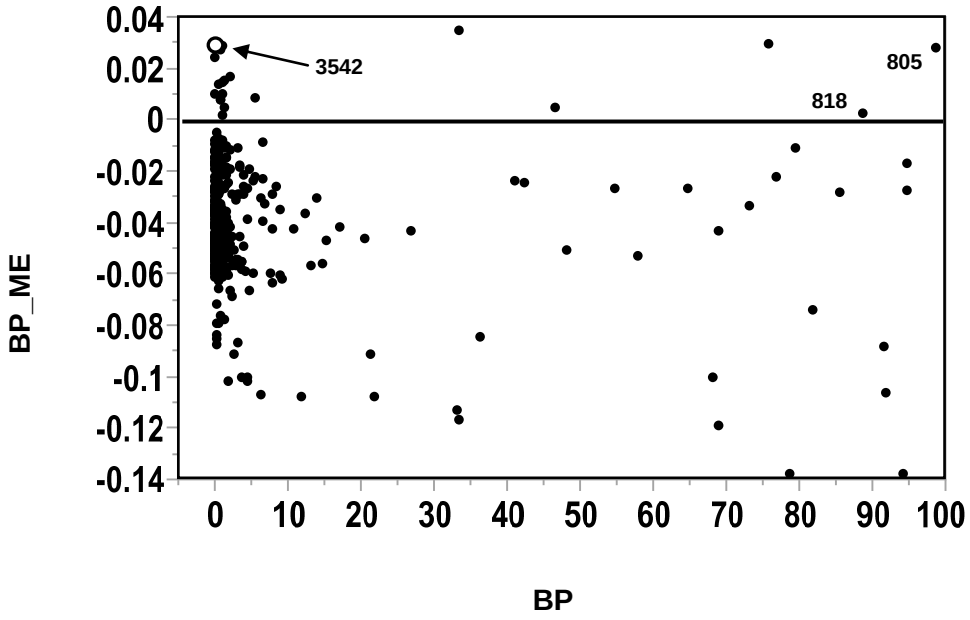


Figure 5.4. Marginal Effects of BP

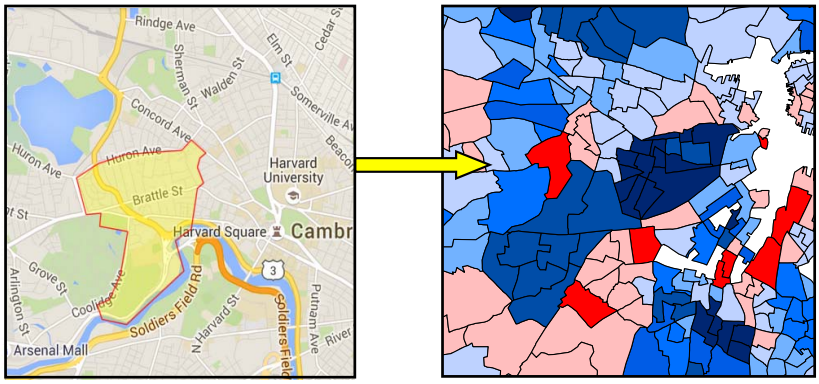
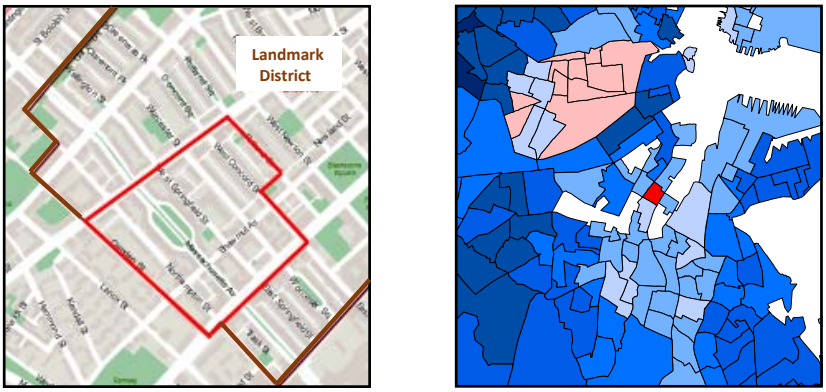


Figure 5.5. Tract 3542 near Harvard Square



(a) LANDMARK DISTRICT **(b) AGE Marginal Effects**

Figure 5.6. Census Track 709

Tables

			SPB	BIC	MAP	SPB	BIC	MAP	SPB	BIC	MAP
Simulation 1	**Relevant**	x_1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
		x_2	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	Irrelevant	x_3	0.050	0.102	**0.008**	**0.000**	0.030	0.001	**0.000**	0.018	**0.000**
		x_4	0.050	0.102	**0.008**	**0.000**	0.034	0.002	**0.000**	0.015	**0.000**
		x_5	0.051	0.142	**0.019**	**0.000**	0.034	0.002	**0.000**	0.014	**0.000**
		n	50	50	50	100	100	100	150	150	150

			SPB	BIC	MAP	SPB	BIC	MAP	SPB	BIC	MAP
Simulation 2	**Relevant**	x_1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
		x_2	0.985	**0.999**	0.996	**1**	**1**	**1**	**1**	**1**	**1**
	Irrelevant	x_3	**0.009**	0.300	0.090	**0.003**	0.169	0.040	**0.003**	0.200	0.050
		x_4	**0.063**	0.367	0.176	**0.002**	0.161	0.044	**0.006**	0.197	0.050
		x_5	**0.013**	0.313	0.093	**0.033**	0.200	0.073	**0.000**	0.154	0.028
		n	50	50	50	100	100	100	150	150	150

Table 4.1. Variable Inclusion Probabilities (VIP)

	Deng		Pace and Gilley	
Variable	β	**t**	β	**t**
CRIM	-0.007	6.513	-0.007	-6.830
RM2	0.009	8.613	0.009	8.390
AGE	0.000	-0.478	-0.002	-3.320
lnDIS	-0.032	-0.203	-0.187	-2.630
TAX	-0.001	-4.364	0.000	-3.510
B	0.000	3.765	0.001	5.990
lnLSTAT	-0.017	-10.063	-0.246	-11.350
NOX2	-0.221	-2.469	-0.369	-2.370
LON			-262.610	-1.380
PTRATIO	-0.014	-3.379	-0.017	-3.090
lnRAD	0.011	5.183	0.073	3.720
INDUS	-0.003	-1.510	-0.001	-0.350
ZN	0.000	2.344	0.001	1.810
LAT			555.950	1.990
CHAS	-0.018	-0.686	-0.012	-0.450
ρ	0.513		0.800	

(a) Spatial Regression Results

	SPB Estimates		MAP Estimates	
Variable	**VIP**	**Avg. ME**	**VIP**	**Avg. ME**
CRIM	**1.000**	-0.053	**1.000**	-0.056
RM	**1.000**	4.917	**1.000**	4.915
AGE	**1.000**	-0.074	**1.000**	-0.075
DIS	**1.000**	-1.168	**1.000**	-1.136
TAX	**1.000**	-0.016	**1.000**	-0.015
BP	**1.000**	-0.036	**1.000**	-0.029
LSTAT	**1.000**	-0.352	**1.000**	-0.347
NOX	**1.000**	-2.972	**1.000**	-2.477
LON	**0.995**	-8.722	**0.998**	-8.084
PTRATIO	0.796	-0.258	**0.987**	-0.304
RAD	0.270	0.029	0.570	0.061
INDUS	0.164	-0.013	0.225	-0.011
ZN	0.003	0.000	0.171	0.000
LAT	0.000	0.000	0.352	0.848
CHAS	0.000	0.000	0.000	0.000
ISO. PROB.	0.000		0.000	

(b) VIP results for VLS

Table 5.1. Variable Relevance comparisons