Modeling spatial variation in the relationships between violence and alcohol and drug use: A comparison of geographically weighted regression and Bayesian spatially varying coefficient process models

David C. Wheeler and Lance A. Waller
Department of Biostatistics, Rollins School of Public Health, Emory University

Abstract

Previous studies consistently indicate measurable local associations between alcohol and drug usage and the incidence of violence, which reveal the importance of spatial analysis in the study of alcohol, drugs, and violence. While studies increasingly incorporate spatial correlation among model residuals to improve model performance, many analyses assume associations that are constant across a study area, an assumption that has come under scrutiny in the statistics and geography literature. In this paper, we compare and contrast two approaches for the estimation of potentially spatially varying regression effects. Specifically, we provide a comparison of a geographically weighted regression (GWR) model with a Bayesian spatially varying coefficient process model, where comparable spatial dependence functions are used. We compare the approaches with respect to conceptual structures, computational implementation, and inferential output. We apply both approaches to model the relationships of violent crime with illegal drug arrest and alcohol sales data from Houston, Texas and compare their results. We also present results of diagnostic tools for
collinearity in the GWR model, where collinearity can lead to correlated regression coefficients that can be problematic for inferential purposes.

Keywords: GWR, Bayesian regression, collinearity, regression diagnostics

Introduction

Statistical models provide both estimation and inference of associations between an outcome of interest and a set of covariates measured on the same experimental units. The statistical theory of linear and generalized linear models provides the widely-used analytic tools of linear regression (for Gaussian outcomes) and binomial or Poisson regression (for binary/percentage and count outcomes), respectively, which permeate almost every scientific discipline. Applying such tools to spatially-referenced data complicates matters, primarily due to violations of assumptions underlying standard regression methodology, namely the independence of observations and the constancy of association between observations and covariates. The first assumption is violated by Tobler’s so called “First Law of Geography” (Tobler 1970), a concept central to the analysis of spatially-referenced data, which suggests that observations taken near each other in space are likely to be more similar (positively correlated) than those taken farther apart. Such correlation may be the result of direct influences between observations (e.g., a high number of infections in one district may result in a high number in a neighboring districts due to interactions between the two populations), or indirect influences of important but unmeasured environmental aspects impacting both the outcome and covariate values (e.g., temperature, humidity, or elevation). The second issue,
outcome-covariate associations that vary across space, is receiving increasing attention in the geography, statistics, and social science literatures and we focus our attention here.

In this paper, we describe and contrast two approaches for estimating spatially-varying associations between outcomes and covariates: geographically weighted regression (Fotheringham et al. 2002) and spatial random effects models (Gelfand et al. 2003). We describe how the two approaches are implemented in different ways, and how each addresses inferential questions regarding spatially-varying associations. We base discussion on the analysis of a particular data set exploring associations between local levels of violence, alcohol sales, and illegal drug activities in Houston, Texas.

**Violence, Drugs, Alcohol, and Space**

To set the stage for our comparison of analytic techniques, note that lay and empirical explanations for local rates of violent crime usually focus upon one of two features of neighborhoods: (1) characteristics of the *people* living in those neighborhoods, or (2) characteristics of the *places* in which they live. More formal crime theories also generally assert that characteristics of people, places, or both people and places are related to observed rates of violence in neighborhood areas (Sampson and Lauritsen 1994). *Social normative* theories and theories of *social disorganization* tend to focus upon population characteristics associated with violence, such as residential mobility, minority status and poverty (Shaw and McKay 1947; Sampson et al. 1997). Conversely, *routine activities* theories focus upon the characteristics of places that are associated with violence, such as areas with abandoned housing or greater retail activity (Felson 1987; Felson et al. 1997). The dividing lines
between people and place characteristics are, of course, somewhat arbitrary; most theorists incorporate both characteristics in their explanations of violence and assert that the interaction between people in places underlies social supports for violent acts (Garofalo 1987; Greenberg and Schneider 1994); this interaction is precisely what we seek to explore, quantify, and model in spatial analyses of related data.

Two recent articles focus upon the integration of theories suggesting mechanisms by which characteristics of places (typically from the perspective of routine activities theory), and aspects of populations (typically from the perspective of social disorganization theory) affect crime rates across community areas. Smith et al. (2000) note that both population and place characteristics may differentially affect crime rates across community areas, while Rice and Smith (2002) stress that both social normative and routine activity concepts are necessary for the development of adequate ecological models of crime. Importantly, these contributions indicate that spatial diffusion may characterize the distribution of crime rates around certain geographic areas of communities, with crime rates in one area affecting crime rates in nearby areas. Although these studies indicate that some spatial processes characterize some aspects of the spatial distribution of crime rates, they do not test whether spatial interactions contribute to crime rates independent of effects related to local population and place characteristics. In other words, they do not answer the question asking whether exogenous characteristics of populations and places, measured in different geographic locations, interact to affect crime rates. Extending this further, our motivating example explores whether such exogenous characteristics impact violence differently in different geographic locations.
More specifically, we wish to link the spatial distribution of alcohol outlets to the general theoretical concept of crime potentials (i.e., the likelihood that crime will be exhibited in an area as a function of people or place characteristics; Brantingham and Brantingham 1993, 1999). Environmental criminologists have been concerned with features of the environment that lead to greater or lesser crime rates and, like the rest of the field, observe that it is a combination of people and place characteristics that underlie crime rates. Since individual actors are ultimately responsible for crime in their communities, crime potentials related to population characteristics are important and represent the production of crime from people living in community areas. Place characteristics also are recognized as facilitating crime by focusing human activities in certain places where social control is weak and social interactions that lead to crime are more likely.

An important aspect of crime potential theory is its claim that unique spatial dynamics underlie the geographic distributions of crimes across neighborhood areas. In the case of interpersonal violence, individuals likely to commit violent acts may interact with others in locations where violence is enabled by place characteristics. Thus, three relevant questions are suggested by this theory:

1. What population characteristics associate with increased incidence of violence?
2. In what places is violence more likely to be exhibited?
3. What spatial interactions link people and place characteristics to violence?

The first question has been addressed in considerable depth by criminologists (see Sampson and Lauritsen 1994). Preliminary answers to the second question have been suggested by routine activity theorists (Felson et al. 1997). Provisional answers to the third question have
been suggested by environmental criminologists (Brantingham and Brantingham 1993, 1999) and provide the motivation for the empirical investigation below.

Focusing on our motivating example, the rates of interpersonal violence appear to increase in and around locations of alcohol outlets, particularly bars and taverns, putatively because (1) bars are often places that attract clientele likely to be involved in violent interactions (e.g., young males), (2) bars are often located in community areas with less guardianship (e.g., retail areas), (3) bars provide opportunities for social interactions that may lead to violence, and (4) bars provide an intoxicating substance that appears to disinhibit aggression among males (Pihl et al. 1997). Indeed, across repeated empirical studies the locations of bars and taverns (Roncek and Maier 1991; Scribner et al. 1995; Gorman et al. 2001; Lipton and Gruenewald 2002) and sales through alcohol outlets (Stevenson et al. 1999) have been correlated with higher rates of violence. Similar arguments have been put forward to support empirically observed cross-sectional relationships between rates of violence and locations of off-premise establishments (Scribner et al. 1995; Alaniz et al. 1998). These effects, however, may be related to other criminogenic aspects of the environments of off-premise outlets (e.g., illegal drug activity and prostitution; Alaniz et al. 1998). There also remains some question as to whether the effects related to alcohol outlets are due to the outlets themselves or are related to general characteristics of areas in which alcohol outlets are located.

**Houston Data**
Building on the conceptual structure above and to motivate and illustrate the statistical methods below, we integrate three datasets pertaining to 439 census tracts within the City of Houston, Texas (based on the 2000 US Census) in the present analysis. The first dataset used in the study includes initial monthly reports of the four violent crime categories (i.e., murder, rape, robbery and aggregated assault) for the City of Houston. We extracted these data from the city police department website. Approximately 98% of the violent crime locations were successfully geocoded and aggregated to the census tract level. The violent crime data are based on first reports of offenses (that is, before investigation and final classification of crimes). Such “call for assistance” data have been used in some previous studies of alcohol availability and violent crime, and have strengths as well as limitations relative to official crime records (Nelson et al. 2001). However, given that data pertaining to officially reported offences have been used in most previous studies, Gorman et al. (2005) conducted a comparison of the total reports of violent crime contained on the police department website for the year 2000 with the official total contained in the 2000 Uniform Crime Reports. This report showed modest discrepancies between the two data sources (under 5% for murder, robbery and aggregated assault, and 10% for race; see Gorman et al. 2005 for details), and therefore all four categories of violent crime were included in the present analysis.

The second dataset describes the location of alcohol outlets active in the city during the year 2000. We accessed this list from the website of the Texas Alcoholic Beverage Commission, and it contained a total of 6,609 outlets, each of which included the name, geographic location, and type of permit or license of the outlet. Almost all of the outlets (99.5%) were successfully geocoded by street address using Centrus Desktop (Group 1
Software 2003). The classification system used by the Texas Alcoholic Beverage Commission allowed differentiation of licenses associated with sales for off-premise and on-premise alcohol consumption. Of the 6,609 outlets in Houston, 1,480 (22.4%) were on-sale, 3,094 (46.8%) off-sale, and 2,035 (30.8%) combined on-/off-sale.

The third dataset used was the Houston police department monthly reports of drug-law violations, accessed from the same website as the violent crime data. A total of 9,985 such crimes is reported for the year 2000, nearly all of which (98%) were successfully geocoded and aggregated to the census tract level in this analysis.

In the analyses below, the violent crime rate represents the outcome variable of interest. The census tract-specific total alcohol sales and drug-crime violations represent our covariates. Since the distribution of violent crime rate is skewed, we use the natural logarithm of it as our outcome variable in models described later. We also use the natural logarithm of each covariate for analysis to maintain linear relationships with the outcome variable and then explore transformed linear associations between violence rates and these adjusted covariates.

Figure 1 displays the violent crime rate by census tracts. The data reveal that some census tracts experience high rates of violence, in some cases the number of violent crime reports exceeds the population size (e.g., the high violence rate tract on the western edge of the study area contains 40 violent crime reports but only 3 residents). These are predominantly low-population-size tracts. Such examples highlight an important complication in analyses such as ours, namely, that violent crime reports from a tract do not necessarily originate from residents of that tract, and the use of census data as a measure of the “at risk” population will be inappropriate in such cases.
Overview of Spatially Varying Coefficient Models

Geographically Weighted Regression (GWR)

In GWR, the data are usually mean measures of aggregate data at fixed points with spatial coordinates; for example, see the Jiangsu province data in Huang and Leung (2002) and the numerous examples in Fotheringham et al. (2002). The spatial coordinates of the data are used to calculate distances that are used in a kernel function to determine weights of spatial dependence between observations. Typically, a separate regression model is fitted at each point location in the dataset, called model calibration locations. For each calibration location, \( s = 1, \ldots, n \), the GWR model at location \( s \) is

\[
y(s) = X(s)\beta(s) + \epsilon(s),
\]

where \( y(s) \) is the dependent variable at location \( s \), \( \beta(s) \) is the column vector of regression coefficients at location \( s \), \( X(s) \) is the row vector of explanatory variables at location \( s \), and \( \epsilon(s) \) is the random error at location \( s \). The estimated regression coefficients at location \( s \) is

\[
\hat{\beta}(s) = (X^T \cdot W(s) \cdot X)^{-1} X^T \cdot W(s) \cdot y,
\]

where \( X = [X(1); X(2); \ldots; X(n)]^T \) is the design matrix of explanatory variables, which typically includes a column of 1’s for the intercept; \( W(s) = diag[w_1(s), \ldots, w_n(s)] \) is the
diagonal weights matrix that is calculated for each calibration location $s$; $y$ is the $n \times 1$ vector of dependent variables; and $\hat{\beta}(s) = (\hat{\beta}_{s0}, \hat{\beta}_{s1}, \ldots, \hat{\beta}_{sp})^T$ is the vector of $p + 1$ local regression coefficients at location $s$ for $p$ explanatory variables and an intercept.

The weights matrix, $W(s)$, is calculated from a kernel function that places more weight on observations that are closer to the calibration location $s$. There are numerous choices for the kernel function, including the bi-square nearest neighbor function, the exponential function, and the Gaussian function. We use the exponential function for the kernel in this paper to match the spatial dependence function used in the SVCP model. The weights from the exponential kernel function are calculated as

$$ w_j(s) = \exp(-d_{sj} / \gamma), $$  

where $d_{sj}$ is the distance between the calibration location $s$ and location $j$, and $\gamma$ is the kernel bandwidth parameter.

To fit the GWR model, the kernel bandwidth is first estimated by cross-validation across all the calibration locations. Cross-validation is an iterative process that finds the kernel bandwidth with the lowest prediction error of all the $y(s)$. For each location $s$, it removes data for observation $s$ in the model calibration at location $s$ and predicts the response $y(s)$ using the other data points and the kernel weights associated with the current bandwidth. Next, the weights are calculated at each calibration location using the kernel function. Finally, the regression coefficients are estimated at each calibration location along with the response estimates by the expression $\hat{y}(s) = X(s)\hat{\beta}(s)$. 

10
To measure the correlation in the sets of estimated regression coefficients, one can calculate their correlation coefficients using

$$C_{sl} = \text{Corr}\left(\{\hat{\beta}_{sk}, \ldots, \hat{\beta}_{sk}\}, \{\hat{\beta}_{sl}, \ldots, \hat{\beta}_{sl}\}\right) = \frac{\sum_{s=1}^{n} (\hat{\beta}_{sk} - \bar{\beta}_{k}) \cdot (\hat{\beta}_{sl} - \bar{\beta}_{l})}{\sqrt{\sum_{s=1}^{n} (\hat{\beta}_{sk} - \bar{\beta}_{k})^2 \cdot \sum_{s=1}^{n} (\hat{\beta}_{sl} - \bar{\beta}_{l})^2}}, \quad (4)$$

where $\bar{\beta}_{k} = \frac{1}{n} \sum_{s=1}^{n} \hat{\beta}_{sk}$. This correlation is subsequently called the overall correlation coefficient of two sets of local regression coefficients and is indicated $C_{sl}$ for the correlation between variables $k$ and $l$ over all locations in the study area. Fotheringham et al. (2002) also present an equation for calculating what they refer to as the local regression coefficient covariance at each location. Technically, this equation is not correct because their version of GWR is not a formal statistical model with kernel weights that are part of the errors. We will, however, use this expression as an exploratory tool for correlation in the local coefficients.

The equation is

$$\text{Cov}[\hat{\beta}(s)] = \sigma^2 \cdot [X^T \cdot W(s) \cdot X]^{-1} \cdot X^T \cdot W^2(s) \cdot X \cdot [X^T \cdot W(s) \cdot X]^{-1}. \quad (5)$$

A local parameter correlation matrix can be calculated from the local covariance matrix as

$$C(s) = \text{diag}^{-1/2}\{\text{Cov}[\hat{\beta}(s)]\} \cdot \text{Cov}[\hat{\beta}(s)] \cdot \text{diag}^{-1/2}\{\text{Cov}[\hat{\beta}(s)]\}, \quad (6)$$
where \( \text{diag}\{\cdot\} \) extracts the diagonal from the covariance matrix. These two equations used for the local coefficient covariance and correlation are only approximate equations because the kernel weights are calculated from the data using cross-validation before the regression coefficients are estimated from the data. The weights are inherently a function of \( y \), as are the regression coefficients, and the correct expression for the coefficient covariance would be non-linear. We will not attempt to derive the exact covariance formula here and instead will use the approximate formula throughout the discussion. The correlation between coefficients for variables \( k \) and \( l \) at location \( s \) is indicated \( C_{kl}s \) and comes from the \((k, l)\) element of the \( C(s) \) correlation matrix. Subsequently, we refer to these correlations as the \textit{local coefficient correlations} at model calibration location \( s \). The pair-wise coefficient correlations can be mapped at each calibration location for each pair of estimated coefficients.

\textit{Bayesian Spatially Varying Coefficient Process (SVCP) Model}

The Bayesian SVCP regression model is specified in a hierarchical manner. The distribution of the data is specified conditional on unknown parameters, whose distribution is in turn specified conditional on other parameters. Following Gelfand et al. (2003), the SVCP model is

\[
\begin{bmatrix} Y \mid \beta, \tau^2 \end{bmatrix} = N(X^T \beta, \tau^2 I),
\]  

(7)
where the bracket notation \([A \mid B]\) denotes the distribution of \(A\) conditional on \(B\). \(Y\) is a vector of responses assumed to be Gaussian conditional on the parameters \(\beta\) and \(\tau^2\); \(\beta\) is a \(np \times 1\) vector of regression coefficient parameters; and \(X^T\) is the \(n \times np\) block diagonal matrix of covariates where each row contains a row from the \(n \times p\) design matrix \(X^*\), along with zeros in the appropriate places (the covariates from \(X^*\) are shifted \(p\) places in each subsequent row in \(X^T\)); \(I\) is the \(n \times n\) identity matrix; and \(\tau^2\) is the error variance.

In the second stage of the hierarchical model, the prior distribution for the regression coefficient parameters is specified as

\[
\left[ \beta \mid \mu_p, \Sigma_p \right] = N(1_{nx1} \otimes \mu_p, \Sigma_p). \tag{8}
\]

The vector \(\mu_p = (\mu_{\beta_1}, \ldots, \mu_{\beta_p})^T\) contains the means of the regression coefficients corresponding to each of the \(p\) explanatory variables. The symbol \(\otimes\) denotes the Kronecker product operator, which is the multiplication of every element in \(1_{nx1}\) by \(\mu_p\). The prior on the regression coefficients in the SVCP model takes into account the possible spatial dependence in the coefficients through the covariance, \(\Sigma_p\). For \(\beta_p = [\beta_{p1}, \ldots, \beta_{pn}]\), we can assume a priori that each \(\beta_p\) follows an areal unit model (e.g., the conditional autoregressive model or simultaneous autoregressive model; see Banerjee et al. 2004) or specify the prior on \(\beta_p\) using a geostatistical approach, where a parametric distance-based covariance function is specified. We focus on a geostatistical prior specification of the regression coefficients and assume an exponential spatial dependence function. The prior covariance matrix for the \(p\) different
types of $\beta$'s at each of $n$ locations, $\Sigma_{\beta}$, can have either a separable or nonseparable form. The separable form has two distinct components, one for the spatial dependence in the regression coefficients and one for the within site dependence between coefficients of the same type. Following Gelfand et al. (2003), we assume a separable covariance matrix for $\beta$ of the form

$$\Sigma_{\beta} = H(\phi) \otimes T,$$  

(9)

where $T$ is a positive-definite $p \times p$ matrix for the covariance of the regression coefficients at any spatial location, $H(\phi)$ is the $n \times n$ correlation matrix that captures the spatial association between the $n$ locations, and $\phi$ is an unknown spatial dependence parameter. In the prior specification for $\beta$ (equation 9), the Kronecker product results in a $np \times np$ positive definite covariance matrix, since $H(\phi)$ and $T$ are both positive definite. The elements of the correlation matrix $H(\phi)$, $H(\phi)_{ij} = \rho(s_i - s_j; \phi)$, are calculated from the exponential function $\rho(d; \phi) = \exp(-d^2 \phi)$.

With the separable cross-covariance function, each of the $p$ coefficients represented in the covariance is assumed to have the same spatial dependence structure. This aligns with the assumption of equal spatial ranges for each explanatory variable in the GWR model, and justifies our use of the separable form. The separable cross-covariance form also has the property that the covariance between $\beta$'s of the same type is constant across space. While a separable assumption is restrictive, one advantage to the separable covariance is that it is more convenient computationally than a nonseparable one and reduces the number of
operations needed for matrix inversion in simulating from the posterior distribution of the parameters. A nonseparable form would allow for distinct spatial ranges for each coefficient process, and is therefore more flexible. An example of the use of a nonseparable covariance matrix in a Bayesian regression model is Banerjee and Johnson (2005), who use a linear model of coregionalization to specify the prior on \( \beta \) (see also Banerjee et al. (2004) and Gelfand et al. (2004) for discussions of models of coregionalization).

The specification of the Bayesian SVCP model in equations (7), (8), and (9) is complete with the specification of the prior distributions of the parameters. The prior for the error variance is inverse gamma with hyperparameters \( a \) and \( b \), \( \tau^2 \sim IG(a,b) \). Hyperparameters control the mean and variance of the prior distribution of the parameter. The prior for the coefficient means is normal with hyperparameters \( \mu \) and \( \sigma^2 \), \( \mu_\beta \sim N(\mu, \sigma^2 I) \). The prior for the covariance matrix \( T \) is inverse Wishart with hyperparameters \( v \) and \( \Omega \), \( T \sim IW_v(\Omega^{-1}) \). These priors are conjugate priors, and are used for computational convenience. The prior for the spatial dependence parameter \( \phi \) is gamma with hyperparameters \( \alpha \) and \( \lambda \), \( \phi \sim G(\alpha, \lambda) \).

Inference on the parameters in the SVCP model is based on the posterior distribution \( \theta | y \) of the parameters \( \theta = (\beta, \tau^2, \mu_\beta, \phi, T) \), which can be obtained using Bayes Theorem:

\[
\theta | y \propto [y | \theta][\theta].
\] (10)

In other words, the posterior distribution for the parameters \( \theta \), conditional on the data, is proportional to the likelihood of the data \( [y | \theta] \), also written as \( p(y | \theta) \), and the prior \( [\theta] \),
also written as $p(\theta)$, for all the parameters. In most situations, it is usually not possible to find an analytic solution for the posterior distribution in complex Bayesian models. Instead, it is common in Bayesian statistics to use simulation-based inference tools such as Markov chain Monte Carlo (MCMC) methods to sample from the posterior distribution of the parameter and base inferences on these samples. MCMC algorithms simulate a Markov chain that has for its stationary distribution the target posterior distribution. The algorithm is run for a sufficient time so that, after a number of ‘burn-in’ iterations, the algorithm converges and the sample path of the Markov chain can be taken to be samples from the posterior distribution of the unknown parameters. The samples from the chain after the ‘burn-in’ are used to summarize inferences on the unknown parameters, where the sample mean or median is typically used as a point estimate of the parameter. In order to check for convergence, it is common to run multiple MCMC algorithms with different starting values, where each is called a chain, and inspect that the sampled posterior distributions are the same for the different chains. Another method to evaluate convergence is to use Gelman’s scale reduction statistic, $\hat{R}$, which has values near 1 for each parameter if the algorithm has converged (Gelman et al. 2004).

In fitting Bayesian models, MCMC algorithms are typically based on the Gibbs sampler (e.g. Casella and George 1992), which iteratively samples from the full conditional distribution for each parameter, conditioning on the current value of the other parameters. The full conditional distribution is the distribution for a parameter given the other parameters in the model. At iteration $j$, the Gibbs sampler for the SVCP model would simulate successively from the following full conditional distributions:
\[
\phi(j) \sim [\phi \mid \beta(j-1), \mu_p(j-1), T(j-1), \tau^2(j-1), Y]
\]
\[
T(j) \sim [T \mid \beta(j-1), \mu_p(j-1), \phi(j), \tau^2(j-1), Y]
\]
\[
\tau^2(j) \sim [\tau^2 \mid \beta(j-1), \mu_p(j-1), T(j), \phi(j), Y]
\]
\[
\mu_p(j) \sim [\mu_p \mid \beta(j-1), \phi(j), T(j), \tau^2(j), Y]
\]
\[
\beta(j) \sim [\beta \mid \phi(j), \mu_p(j), T(j), \tau^2(j), Y].
\]

Clearly, to use the Gibbs sampler, one must be able to derive the full conditional distribution for each parameter. The full conditional distribution for each parameter is derived by taking the product of the appropriate likelihood function and the priors for all the parameters and then simplifying the expression for the parameter of interest by ignoring terms that do not include the parameter of interest. Unlike the other parameters, the full conditional distribution of \( \phi \) cannot be found in closed form.

When a conditional distribution for a parameter can only be calculated up to a normalizing constant, as is the case with \( \phi \), a Metropolis-Hastings (M-H) step can be used to draw a sample from the full conditional distribution of a parameter (e.g. Chib and Greenberg 1995). In the M-H step, we use a normal random walk proposal density that is centered on the current value of \( \phi \) and has a variance \( s^2 \) that is tuned to produce an adequate acceptance rate of the proposed value of \( \phi \). The proposed value of \( \phi \) is accepted if the ratio of the unnormalized full conditional distribution with the proposed value over the unnormalized full conditional distribution with the current value of \( \phi \) is greater than 1 or, if not, is greater than a randomly drawn uniform variable with a range of (0,1).

A feature of the SVCP model is that the correlation between explanatory variable coefficients is explicitly modeled. With the separable covariance matrix, the posterior
correlation of the regression coefficients $k$ and $l$ across all locations is $T_{kl}/\sqrt{T_{kk}T_{ll}}$. Another expression for this correlation is

$$\frac{T_{kl}}{\sqrt{T_{kk}T_{ll}}} = \frac{\text{cov}(\beta_k(s), \beta_l(s+h))}{\sqrt{\text{cov}(\beta_k(s), \beta_k(s+h))\text{cov}(\beta_l(s), \beta_l(s+h))}},$$

(12)

which does not depend on $h$, hence, the correlation of $(\beta_k, \beta_l)$ does not depend on distance $(h)$ between locations and is the same across the study area (Gelfand et al. 2004). We make use of this generalized coefficient correlation in the next section when evaluating dependence in the coefficients in the simulation studies.

**Results**

For the Houston crime data, the GWR model is

$$y(i) = \beta_0(i) + \beta_1(i)x_1(i) + \beta_2(i)x_2(i) + \varepsilon(i),$$

(13)

where $y$ is the natural log of the number of violent crimes (murder, robbery, rape, and aggregated assault) per 100 persons, $x_1$ is the natural log of the number of drug law violations per 100 persons, $x_2$ is the natural log of the number of alcohol outlets per 100 persons, and $i$ is the index for census tracts.

Through cross-validation, the estimated GWR kernel bandwidth $\gamma = 0.89$. The root mean square prediction error (RMSPE) associated with this estimated bandwidth is 0.720.
The root mean square error (RMSE) of the response variable for the GWR model with this estimated bandwidth and regression coefficients estimated with the bandwidth is 0.342.

For the SVCP model, the prior specification for the model is as follows. We use a vague normal, $N(0, 10^4 \mathbf{I})$, for $\mu$, a vague inverse Wishart, $IW(p+1, 0.1 \mathbf{I})$, for $\mathbf{T}$, and a vague inverse gamma, $IG(1, 0.01)$, for $\tau^2$, where $\mathbf{I}$ is the identity matrix of dimension $p$. For the spatial dependence parameter $\phi$, we use a gamma, $G(0.07, .01)$, which has a mean of 7 and variance of 700. Random starting values are used for the parameter Markov chains.

Through MCMC sampling of the parameter joint posterior distribution, the spatial dependence parameter in the Bayesian SVCP model is estimated as $\hat{\phi} = 0.732$ using the posterior mean. A sample size of 2,000 was used in the MCMC, with a “burn-in” of 4,000 iterations. The posterior mean estimate of the error variance is $\hat{\tau}^2 = 0.0197$ and the posterior mean estimates for the regression coefficient means are $\hat{\mu}_{\beta} = (0.676, 0.572, 0.110)^T$. The posterior mean estimate of the within variance matrix for the regression coefficients is

$$
\hat{\mathbf{T}} = \begin{bmatrix} 0.230 & -0.049 & 0.078 \\ 0.084 & -0.027 & 0.042 \end{bmatrix}.
$$

The RMSE for the response variable is 0.040 with the SVCP model. The estimates for the response variable, $\hat{y}$, are plotted against $y$ for the GWR and SVCP models in Figure 2. The SVCP model fits the response variable considerably better than GWR, and has a RMSE approximately one order of magnitude below that of GWR.

Histograms of the estimated regression coefficients from the GWR and SVCP models are plotted in Figure 3 (intercept), Figure 4 (drug), and Figure 5 (alcohol). The GWR estimated regression coefficients show more variation than do the SVCP model coefficients, especially for the intercept and alcohol, as evident from the ranges in the histograms. This
suggests the SVCP model is shrinking the regression coefficients to their respective means better than GWR. The extra variation in the GWR coefficients is also evident in scatter plots of $\hat{\beta}_1$ versus $\hat{\beta}_0$ (Figure 6), $\hat{\beta}_2$ versus $\hat{\beta}_0$ (Figure 7), and $\hat{\beta}_2$ versus $\hat{\beta}_1$ (Figure 8) for the GWR and SVCP models. Scatter plots like these have been suggested by Wheeler and Tiefelsdorf (2005) to check on the level of dependence in estimated GWR coefficients. Indeed, it appears there is substantial dependence in the regression coefficients for both models, but particularly for the SVCP model. The overall correlations for the GWR coefficients are $C_{01} = 0.23$, $C_{02} = 0.59$, $C_{12} = -0.41$ and the overall correlations for the SVCP coefficients are $C_{01} = -0.74$, $C_{02} = 0.87$, $C_{12} = -0.73$. The within area correlations between regression coefficients for the SVCP are $T_{01}/\sqrt{T_{00}T_{11}} = -0.35$, $T_{02}/\sqrt{T_{00}T_{22}} = 0.79$, and $T_{12}/\sqrt{T_{11}T_{22}} = -0.46$. The estimated regression coefficients are mapped in Figure 9 for GWR and in Figure 10 for the SVCP model. The figures show that the GWR estimated regression coefficients are spatially smoother than the SVCP estimated regression coefficients. This is in agreement with results from previous analysis of the Houston crime data by Waller et al (2007), where GWR coefficients were smoother than those from a Bayesian spatially varying coefficient model that used conditionally autoregressive spatial dependent specification for the regression coefficients. There is also somewhat of a pattern of complementarity in the estimated coefficients for alcohol outlets and drug violations, where in some peripheral areas the alcohol coefficient is high and the drug violations coefficient is low. This pattern is stronger in the GWR coefficients than in the SVCP model coefficients. This suggests that the complementary pattern in spatially varying coefficients is controlled.
less by correlation in the coefficients and perhaps more by inflated variance of the coefficients.

Collinearity Diagnostics

To further explore dependence in the regression coefficients, we apply a collinearity diagnostic for the GWR model and explore sensitivity to the prior specifications in the SVCP model. To assess collinearity in the GWR model, we use the variance-decomposition proportion and condition index diagnostic tool introduced by Wheeler (2007). The variance-decomposition is done through singular value decomposition and it has an associated condition index, which is the ratio of the largest singular value to the smallest singular value. In diagnosing collinearity, the larger the condition index, the stronger is the collinearity among the columns of the GWR weighted design matrix. Belsley (1991) recommends a conservative value of 30 for a condition index that indicates collinearity, but suggests the threshold value could be as low as 10 when there are large variance proportions for the same component. The variance-decomposition proportion is the proportion of the variance of a regression coefficient that is affiliated with one component of its decomposition. In addition, the presence of two or more variance proportions greater than 0.5 in one component of the variance-decomposition indicates that collinearity exists between at least two regression terms, one of which may be the intercept.

The variance-decomposition proportions and condition indexes are listed in Table 1 for records with the largest condition indexes. The GWR estimated bandwidth is used in the variance-decomposition of the kernel weighted design matrix to assess the collinearity in the
model. These 10 records are labeled in the left plot of estimated GWR coefficients for the drug and alcohol covariates in Figure 8. These labeled records comprise many of the more extreme points in the plot. Observation 153 is clearly the most extreme of the points, as it has the largest value for the drug rate effect and the smallest value for the alcohol rate effect. In Table 1, this record has large variance proportions for the same component for all three regression terms. Of the 439 records in the dataset, 5 have a condition index above 30, 10 have a condition index above 20, and 41 have a condition index above 10. There are 411 records in the data with large variance proportions (> 0.5) from the same component, with the shared component being between a covariate and the intercept for some records and between the two covariates for other records. Overall, the variance-decomposition proportions and condition index values indicate the presence of some local collinearity in the GWR model. Incidentally, the only one of the more extreme GWR observations in Table 1 that is clearly visible in the plot for the SVCP coefficients in Figure 8 is observation 153. The other nine observations listed in the table have been shrunken towards the mean values in the SVCP model.

SVCP Model and Coefficient Shrinkage

It is not completely unexpected that the SVCP model produces a more compact set of regression coefficients than GWR given the discussions in the literature of ridge regression solutions as Bayes estimates. Hoerl and Kennard (1970) first introduced ridge regression to overcome ill conditioned design matrices. Ridge regression coefficients minimize the residual sum of squares along with a penalty on the size of the squared coefficients as
\[ \hat{\beta}^R = \arg \min_{\beta} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{k=1}^{p} x_{ik} \beta_k \right)^2 + \lambda \sum_{k=1}^{p} \beta_k^2 \right\}, \]  

(14)

where \( \lambda \) is the ridge regression parameter that controls the amount of shrinkage in the regression coefficients (Hastie et al. 2001). Works by Lindley and Smith (1972) and Goldstein (1976) show that ridge regression coefficient estimates may be viewed as Bayesian regression coefficient posterior means under specific vague priors. Hastie et al. (2001) also describe the ridge regression solutions as Bayes estimates, where ridge regression uses independent normal distributions for each coefficient \( \beta_k \) prior. If the prior for each regression coefficient \( \beta_k \) is \( N(0, \sigma^2) \), independent of the others, then the negative log posterior density of the regression coefficients \( \beta \) is equal to the expression in the braces in the ridge regression coefficient equation (14), with \( \lambda = \tau^2 / \sigma^2 \), where \( \tau^2 \) is the error variance. Specifically, in the Bayesian regression model with \( y \sim N(X\beta, \tau^2 I) \) and an independent prior \( \beta \sim N(0, \sigma^2) \) for each coefficient, the posterior for the coefficients can be expressed as

\[
[\beta | \tau^2, \sigma^2, y] \propto \exp \left( -\frac{1}{2\tau^2} \sum_{i=1}^{n} (y_i - \sum_{k=1}^{p} x_{ik} \beta_k)^2 \right) \exp \left( -\frac{1}{2\sigma^2} \sum_{k=1}^{p} (\beta_k - 0)^2 \right),
\]  

(15)

where for convenience of notation the variables have been centered. The negative log posterior density of \( \beta \) is then found through algebra to be
with the ridge shrinkage parameter \( \lambda = \frac{r^2}{\sigma^2} \). This illustrates that the ridge regression estimate is the mean of the posterior distribution with a Gaussian prior and Gaussian data model, and that the ridge shrinkage parameter is a ratio of the error variance and common regression coefficient variance.

The view of ridge regression solutions as Bayes estimates suggests that the Bayesian SVCP model coefficients can be viewed as ridge regression estimates because of the normal distribution prior for the regression coefficients in the SVCP model. Granted, the prior in the SVCP is more complicated than the independent normal prior in the traditional Bayesian regression model due to the spatial component in the covariance matrix. The posterior distribution for the coefficients of the SVCP model can be expressed with convenient notation as

\[
[\beta \mid \tau^2, \mu_p, \Sigma_p; y] \propto \exp \left( -\frac{1}{2} (y - X\beta)^T (\tau^2 I)^{-1} (y - X\beta) \right) \\
\times \exp \left( -\frac{1}{2} (\beta - (1 \otimes \mu_p))^T \Sigma_p^{-1} (\beta - (1 \otimes \mu_p)) \right) \\
\propto \exp \left( -\frac{1}{2} [(y - X\beta)^T (y - X\beta) + (\beta - (1 \otimes \mu_p))^T \tau^2 (H(\phi) \otimes T)^{-1} (\beta - (1 \otimes \mu_p))] \right).
\]

(17)

The negative log posterior density of \( \beta \) up to a constant is then
\[(y - X\beta)^T (y - X\beta) + (\beta - (I \otimes \mu_p))^T \tau^2 (H(\phi) \otimes T)^{-1} (\beta - (I \otimes \mu_p)), \] (18)

where the shrinkage term is unconventionally a matrix \( \Lambda \) that is calculated as
\[\tau^2 \cdot H^{-1}(\phi) \otimes T^{-1}.\] Therefore, the amount of shrinkage on \( \beta \) towards the mean \( \mu_\beta \) depends on \( \tau^2, \phi, \) and \( T \) in the SVCP model. Keeping \( \tau^2 \) and \( \phi \) constant, the shrinkage factor is controlled by the covariance matrix \( T \). In \( T \), the terms that control the spatial variation in the regression coefficients are the variances, i.e. the diagonal terms in the matrix, \( T_{11}, T_{22}, \ldots, T_{pp} \). If \( \tau^2 \) and \( \phi \) are held constant and the variances are increased in the diagonal of \( T \), then the shrinkage factor will decrease and coefficients may vary more from their respective means. Conversely, if the variances in \( T \) are decreased while holding \( \tau^2 \) and \( \phi \) steady, then the shrinkage factor will increase and the regression coefficients may shrink more to their respective means. In the Bayesian framework, the regression coefficient variances in \( T \) may be effectively increased or decreased through its prior.

Returning to the SVCP regression coefficients for violent crime in Houston, the level of overall correlation between the estimated coefficients for drug violations and alcohol outlets is somewhat surprising given the view of the Bayesian SVCP coefficients as shrinkage estimates. However, it is possible to control the amount of correlation in the estimated SVCP regression coefficients by selecting a prior for the within area covariance matrix that favors no correlation between the drug violations and alcohol terms. In practice, this is done by specifying a relatively large scale matrix with off-diagonal elements of 0 in the inverse Wishart prior distribution for \( T \) and increasing the degrees of freedom to be greater than \( p+1 \), where \( p \) is the number of linear regression terms. By increasing the prior
variances, the posterior covariance can be decreased between regression terms. For example, changing the prior for $T$ from the vague $IW(p+1, 0.1\mathbf{I})$ to the more informative $IW(p+2, 10\mathbf{I})$ should have some effect on the posterior of $T$ and therefore the regression coefficients $\hat{\beta}$.

To illustrate this effect of changing the prior, we performed sampling from the joint posterior distribution using MCMC as before, but this time with a prior of $IW(p+2, 10\mathbf{I})$ for $T$. The resulting estimated regression coefficients are mapped in Figure 11. The informative prior results in more smoothed regression coefficients, somewhat similar to those from GWR in Figure 9. Increasing the prior variances of the regression terms allows them to move more away from their respective means and results in more smoothed patterns of estimated coefficients. The scatter plot of the estimated SVCP regression coefficients for alcohol outlets versus drug violations is displayed in Figure 12 for the vague prior on the left and the informative prior on the right. The scatter plot shows that there is less overall correlation in the SVCP coefficients when the informative prior is utilized. In fact the overall correlation decreases to $C_{12} = -0.32$ and the within area correlation decreases to $T_{12}/\sqrt{T_{11}T_{22}} = -0.26$. However, the RMSE of $\hat{y}$ increases to 0.061 with the informative prior.

The results of changing the prior to an informative one with more within-area variance and less covariance shows that one can successfully control the amount of correlation in estimated regression coefficients in the SVCP model rather easily, with the cost of a potential increase in response variable estimation error. In addition, it is possible to get more smoothed coefficients, similar to those from GWR, by changing the within area variance prior.

*Model Prediction*
Also of interest in spatial regression models and violent crime analysis is the prediction of the response variable for a new observation, for example the crime rate at a new census tract. Both GWR and the SVCP can predict values at new observation locations in space. To compare the prediction performance of the GWR and SVCP models, we fitted the two models again using a 90% subset of the data and then predicted the response variable for the remaining hold-out random sample of approximately 10% of the original data observations using the model parameter estimates from the larger data sample. The sampling is done without replacement, so the two samples are mutually exclusive. There are 44 observations (census tracts) in the hold-out sample and 395 in the model calibration sample. The estimated kernel bandwidth for GWR was $\hat{\gamma} = 1.03$. The regression coefficient mean estimates for the hold-out sample distributions were drawn from the posterior prediction distribution for $\hat{\beta}$.

The RMSPE for $\hat{y}$ for the GWR model is 0.518 and the RMSPE for the SVCP model is 0.521, so there is less than a 1% difference in the model prediction errors. The model predictions are plotted in Figure 13, which conveys that the performance of the two models is similar in predicting the hold-out sample response values. The correlation in the predicted and actual responses is 0.92 for GWR and 0.90 for the SVCP model.

One key advantage in prediction of the response variable for the SVCP model is the ability to produce prediction intervals. For each prediction, a prediction interval of a desired confidence level can be generated using draws from the parameter posterior distributions. See, for example, Figure 14, which has 95% prediction intervals drawn for the SVCP model point predictions (plusses) for the response variable along with the GWR predictions.
(triangles) and the actual values (circles). Note that 43 of the 44 (98%) prediction intervals contain the observed violent crime rate. The GWR methodology does not directly produce prediction intervals, and it is not clear how one would generate them through resampling of the data. The fact that one can create prediction intervals readily from the existing MCMC samples is a major advantage for the SVCP model. In fact, the quantification of such uncertainty is an advantage of the Bayesian paradigm in general.

It should be noted that the criterion used in model parameter estimation is not the same for the GWR and SVCP models. GWR uses leave-one-out cross-validation to minimize the prediction error of the data set where each observation is predicted by all other observations. The criterion for kernel bandwidth selection we use is explicitly a function of the response variable prediction. GWR then uses the optimal bandwidth to define the kernel values for the location-specific spatial weights matrix, which in turn defines the parameter estimates. For the Bayesian SVCP model, all inference follows from the joint posterior distribution of model parameters. We sample directly from the joint posterior distribution of the parameters, where values are drawn more frequently that are more likely given the data. More likely parameter values effectively maximize the likelihood of the observed data (assuming an uninformative prior) without explicitly considering prediction of new data. Therefore, it is not surprising that GWR with cross-validation estimation of the kernel bandwidth would perform better than the SVCP model for prediction. It would be interesting in a future study to compare the two models when using an Akaike Information Criterion (AIC), as discussed in Fotheringham et al. (2002), in place of cross-validation in GWR, given that the AIC is not directly based on prediction of the response variable. Also, only one hold-
out sample out of a large number of possible samples was used in this analysis. A more comprehensive comparison would use a larger number of samples.

**Conclusions**

In the preceding sections, we review the basic structure of two approaches for estimating geographic variation in the association between our outcome (local rates of violence crime) to two covariates of interest (local alcohol sales and local illegal drug activities). In summary, geographically weighted regression provides a descriptive approach using geographic weights to focus parameter estimation on nearby data, while spatially-varying coefficient models utilize a random effects structure to define spatial correlations within a probability model to influence estimated associations. We find some qualitative similarity in model estimates, but note that the two approaches specify spatial variation in quite different manners. Overall, we find GWR provides a quicker descriptive result providing maps of smoothed general differences in association, but it is somewhat limited in terms of statistical inference regarding the amount and extent of spatial pattern, especially with respect to allowing different amounts of spatial adjustment for different covariates. In contrast, we find the random effects spatially-varying coefficient models to provide a broader inferential basis for analysis of spatially referenced data that in turn allows a richer interpretation of estimated associations, but at a considerable computational cost. This computational cost is in terms of not only computing time, but also setting up the model in a manner friendly to accurate and efficient MCMC implementation. More methodological research is required to determine the most flexible and robust structures, and the data
required for implementation. In our opinion, the payoff for the extra effort required to set up the random effects model is a framework for model-based inference, and a broader class of possible models to fit within that framework.

In closing, the data analysis presented here illustrates the importance of spatial variation in associations between alcohol distribution and violence, and presents two statistical approaches for estimating such spatial variations. The two approaches take different routes to provide spatially varying effect estimates, and both are non-trivial to fit. The results are intriguing but not complete. Next steps include incorporation of additional local covariates representing both local populations and neighborhoods, offering additional insight into the associations of interest, as well as stretching current methodological and computational capabilities.

The issue of covariate collinearity and parameter correlation also merits additional investigation. Our analysis reveals correlations between model parameters, and in both the GWR and Bayesian SVCP models. The spatially varying coefficient model allows us to model this correlation and, as noted, we find the incorporation of such correlation to improve overall model fit. These inter-parameter correlations are likely due to multiple causes including (but not limited to) collinearity between the two covariates considered in our example. As we demonstrated, it is possible with the SVCP model to control the amount of correlation in the regression coefficients. In addition, the SVCP model naturally shrinks regression coefficients to counter model collinearity. These features, along with readily available prediction intervals for the response variable, offer several advantages of the SVCP model over GWR.
References


Greenberg M, Schneider DF (1994) Violence in American cities: Young black males is the answer, but what was the question. *Social Science and Medicine* 39: 179 - 187

Group 1 Software (2003) *Centrus Desktop, Version 4.01.01*. Group 1 Software; Boulder, CO


Roncek DW, Maier PA (1991) Bars, blocks, and crimes revisited: Linking the theory of routine activities to the empiricism of "hot spots". *Criminology* 29: 725 - 753


### Tables

<table>
<thead>
<tr>
<th>ID</th>
<th>k</th>
<th>p₁</th>
<th>p₂</th>
<th>p₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27.60</td>
<td>0.996</td>
<td>0.995</td>
<td>0.136</td>
</tr>
<tr>
<td>2</td>
<td>87.66</td>
<td>0.992</td>
<td>0.992</td>
<td>0.001</td>
</tr>
<tr>
<td>5</td>
<td>21.29</td>
<td>0.995</td>
<td>0.993</td>
<td>0.188</td>
</tr>
<tr>
<td>27</td>
<td>24.25</td>
<td>0.997</td>
<td>0.690</td>
<td>0.947</td>
</tr>
<tr>
<td>33</td>
<td>35.45</td>
<td>0.865</td>
<td>0.949</td>
<td>0.045</td>
</tr>
<tr>
<td>67</td>
<td>29.49</td>
<td>0.994</td>
<td>0.982</td>
<td>0.371</td>
</tr>
<tr>
<td>114</td>
<td>40.58</td>
<td>0.739</td>
<td>0.988</td>
<td>0.283</td>
</tr>
<tr>
<td>116</td>
<td>39.45</td>
<td>0.579</td>
<td>0.996</td>
<td>0.922</td>
</tr>
<tr>
<td>153</td>
<td>38.38</td>
<td>0.737</td>
<td>0.999</td>
<td>0.609</td>
</tr>
<tr>
<td>158</td>
<td>21.94</td>
<td>0.955</td>
<td>0.942</td>
<td>0.006</td>
</tr>
</tbody>
</table>

Table 1. Record number, condition index (k), and variance-decomposition proportions (p₁ = intercept, p₂ = drug, p₃ = alcohol) for the GWR model of Houston crime
Figure 1. Violent crime rate by census tract in the City of Houston
Figure 2. The estimates of the response variable versus the response variable for the GWR and SVCP models.
Figure 3. GWR and SVCP estimated regression coefficients by census tracts for the intercept
Figure 4. GWR and SVCP estimated regression coefficients by census tracts for drug violations
Figure 5. GWR and SVCP estimated regression coefficients by census tracts for alcohol outlets
Figure 6. GWR and SVCP estimated coefficients for the intercept and drug violations
Figure 7. GWR and SVCP estimated coefficients for the intercept and alcohol outlets
Figure 8. GWR and SVCP estimated coefficients for drug violations and alcohol outlets
Figure 9. GWR estimated coefficients for the intercept, drug violations, and alcohol outlets
Figure 10. Bayesian SVCP estimated coefficients for the intercept, drug violations, and alcohol outlets
Figure 11. Bayesian SVCP estimated coefficients for the intercept, drug violations, and alcohol outlets with an informative prior for $T$. 
Figure 12. SVCP estimated regression coefficients for drug violations and alcohol outlets with a vague prior for $T$ (left) and an informative prior (right).
Figure 13. GWR (left) and SVCP (right) response variable predictions for a hold-out sample
Figure 14. GWR (triangles) and SVCP (plusses) predictions for the response variable (circles) of a hold-out sample with 95% prediction intervals from the SVCP model