Bayesian Variable Selection for Multivariate Spatially-Varying Coefficient Regression: Application to Physical Activity in Pregnancy

Brian J. Reich[a1], Montserrat Fuentes[a], Amy H. Herring[b]

[a] Department of Statistics, North Carolina State University

[b] Department of Biostatistics and Carolina Population Center, University of North

Carolina at Chapel Hill

Institute of Statistical Mimeo Series # 2617

Abstract

Physical activity has well-documented health benefits for cardiovascular fitness and weight control. For pregnant women, the American College of Obstetricians and Gynecologists currently recommends 30 minutes of moderate exercise on most, if not all, days; however, very few pregnant women achieve this level of activity. Epidemiologists, policy makers, and city planners are interested in whether characteristics of the physical environment in which women live and work have influence on physical activity levels during pregnancy. In this paper we study the associations between physical activity and several factors including personal characteristics, meteorological/air quality variables, and neighborhood characteristics in pregnant women in four counties of North Carolina. We simultaneously analyze six types of physical activity and investigate cross-dependencies between these activity types. Exploratory analysis suggests that the associations are different in different regions. Therefore we use a multivariate regression model with spatially-varying regression coefficients. This model includes a regression parameter for each covariate at each spatial location. For our data with many predictors, some form of dimension reduction is clearly needed. We introduce a Bayesian variable selection procedure to identify subsets of important variables. Our stochastic search algorithm determines the probabilities that each covariate's effect is null, non-null but constant across space, and spatially-varying. As might be expected, we found that individual level covariates had a greater influence on women's activity levels than neighborhood environmental characteristics, and some individual level covariates had spatially-varying associations with the activity levels of pregnant women.

Key words: Physical activity; spatial data; truncated data; Bayesian variable selection; zero-inflation.

1 Introduction

A recent population-based study found that while approximately two-thirds of pregnant women reported some leisure activity during pregnancy, less than one-sixth reported meeting the recommendations for activity (Evenson et al., 2004). Physical activity during pregnancy has well-established benefits, including prevention of excess weight gain (Clapp and Little, 1995) and reduced risk of certain complications such as gestational diabetes (Dempsey et al., 2004) and preeclampsia (Saftlas et al.). During pregnancy, women change their activity patterns and tend to adopt more health promoting behaviors (e. g. avoidance of cigarettes or binge drinking). If healthy behavioral patterns established during pregnancy could be maintained afterwards, a woman's weight status and general health could be improved. Epidemiologists, policy makers, and city planners are interested in whether characteristics of the physical environment in which women live and work have influence on physical activity levels during pregnancy. In this paper we study the associations between six types of moderate to vigorous physical activity and several factors including personal characteristics, meteorological/air quality variables, and neighborhood characteristics in pregnant women in four counties of North Carolina.

Physical activity patterns are highly dependent on geographic location. Our data includes women from Orange, Durham, Alamance, and Chatham counties in central North Carolina, with jittered locations of study subjects plotted in Figure 1a. These counties are quite heterogeneous, with communities ranging from the more urban area of downtown Durham, to the college town of Chapel Hill, to rural areas of Alamance, Chatham, and northern Orange counties. Properly accounting for spatial clustering of the responses is needed to

study the effects of our predictors on physical activity (Reich et al., 2006, Hoeting et al., 2006).

These sources of heterogeneity made investigators suspect that some exposures may have different associations with physical activity in different areas. For example, sidewalks may be more important on busy streets than on quieter ones. Porches on homes in rural areas may be associated with homes on larger lots and residents more likely to spend considerable time in outdoor household physical activity related to gardening or farm upkeep; however, in Chapel Hill, a porch may be a marker of living in a new neighborhood development with an active living design distinguished by homes on smaller lots in an area designed specifically to encourage walking for transportation. To capture these spatially-varying effects, we extend the spatially-varying coefficient model of Gelfand et al. (2003) to the multivariate setting.

Implementing the multivariate spatially-varying coefficients model for these data is challenging for two reasons: (1) the data are clearly non-Gaussian, with a large proportion of zeros corresponding to women with no moderate to vigorous activity of a given type, and (2) there are 30 predictors, so allowing all of the predictors to have spatially-varying effects leads to an over-parameterized model. Figure 1b plots the transformed (log[activity +1]) transportation activity levels. Nearly 80% of the study women reported no transporation-related moderate to vigorous physical activity. One approach would be to dichotomize the data and apply spatial logistic or probit regression to the binary responses (e.g., Kelsall and Diggle, 1998; Paciorek, 2007). While this approach would simplify the analysis, it ignores the large differences in the non-zero responses, e.g., the difference between a woman who rides her bike to work compared to a woman who walks uphill to the bus stop. We elect to retain all the information in the data and simultaneously analyze the zero and continuous

non-zero responses.

Stein (1992) introduces a truncated normal distribution for spatial data. Exploratory analysis shows that this model is not flexible enough to fit the positive observations and have sufficient mass at zero to accommodate the large proportion of zeros. Therefore, we develop a zero-inflated spatially-referenced truncated Gaussian model. Zero-inflated models are common for spatial count data (e.g., Agarwal, 2002) and non-spatial semicontinuous data (Olsen and Schafer, 2001; Berk and Lachenbruch, 2002; Albert and Shen, 2005; Zhang et al., 2006) but have not been applied to semicontinuous spatial data. There is an important distinction between our approach and previous non-spatial models for semicontinuous data. These models include two regression coefficients for each predictor: the first is used to model the probability at zero and the second is used to predict the magnitude of the non-zero responses. Our model includes only a single regression coefficient for each predictor, and thus allows both the zero and non-zero observations to inform about the regression coefficients.

To reduce the dimension of the model we develop a stochastic search variable selection (e.g., George and McCulloch, 1993; Chipman, 1996; George and McCulloch, 1997; Mitchell and Beauchamp; 1998) procedure for spatially-varying coefficients. The literature on variable selection for spatial data is limited. Hoeting et al. (2006) and Huang and Chen (2007) propose classical variable selection methods for univariate Gaussian spatial data. Smith and Fahrmeir (2007) develop a Bayesian variable selection method for univariate data on a lattice with spatially-varying coefficients. Their approach is to perform local variable selection, that is, to allow for a different subset of the covariates to be included at different lattice points. They use an Ising prior to encourage neighboring lattice points to include similar variables. Our approach differs from Smith and Fahrmeir in several ways. Their model is tailored to

4

univariate Gaussian data on a lattice; our data are multivariate, zero-inflated, and not on a lattice. Also, we take a more global view to variable selection in that we select the same variables throughout the spatial domain but allow some of their effects to vary. Smith and Fahrmeir's model does not explicitly place prior probability on variables being completely removed from the model or included throughout the spatial domain with the same effect across space. This is appropriate for their analysis of fMRI data in which the objective is to uncover "hot spots" with a small number of covariates. However, in our setting with many variables, encouraging these models is desirable. Our prior gives non-zero mass to three models for each covariate and for each outcome: (1) no association, (2) spatially non-varying association, and (3) spatially-varying association. This allows us to include efficiently a large number of predictors in the model, without forcing each to have a spatially-varying effect. Our hierarchical model encourages variables to be included/excluded consistently across activity types, producing a more interpretable multivariate model.

The paper proceeds as follows. Section 2 describes the PIN3 physical activity data. Section 3 develops the statistical model, the zero-inflated model and variable selection algorithm. Computational details are given in Section 4. The methods are applied in Section 5 to PIN3 data. Section 6 concludes.

2   Description of the data

The third phase of the Pregnancy, Infection, and Nutrition (PIN3) Study recruited pregnant women seeking prenatal care at the University of North Carolina Hospitals. Because of the cachement area of UNC Hospitals, most study women resided in Orange, Durham, Alamance,

or Chatham counties in central North Carolina. Figure 1a plots the residential locations of the women (random "jitter" has been added to each location to protect the women's identities, which explains why a few observations lie outside the four labeled counties). One aim of the study was to investigate patterns of physical activity during pregnancy. We examine factors related to physical activity during the second trimester of pregnancy in six domains, including leisure or recreational activity, activity done at work, transportation activity, child or adult care activity, and both indoor and outdoor household activities.

Women enrolled in the study were given a one-week recall questionnaire designed to capture moderate and vigorous activity in the past week. Moderate to vigorous activity was defined in the questionnaire as activity that raised the woman's heart rate. The questionnaire assessed both frequency and duration of activity in each of the six domains. Intensity of activity was assessed using published metabolic (MET) tables (Ainsworth, 2000). The MET, 1 kilocalorie per kilogram per hour, is defined as the amount of energy the body uses when sitting quietly. Reported activities were converted to MET equivalents based on the published tables. Examples of MET equivalents for activities reported in the PIN3 study include indoor household activities like vacuuming (3.5 METs), outdoor household activities like raking the lawn (4.3 METs), transportation activities like biking less than 10 mph (4.0 METs), or leisure activities such as a swimming slow freestyle laps (7 METs) or running at a 8.5 minute mile pace (11.5 METs). The number of MET minutes per week in each activity domain was then calculated by summing over the activities within each domain, weighted by the duration of each activity reported. For example, a woman who swam slow laps for one hour twice per week would be assigned $7 \times 2 \times 60$ MET minutes of leisure-time moderate to vigorous activity; a woman who reported biking <10mph to work for 10 minutes each

direction 5 days per week would be assigned $4 \times (10 + 10) \times 5$ MET minutes of moderate to vigorous transportation-related activity.

Potential predictors of physical activity during pregnancy include sociodemographic, meteorological, and neighborhood environmental factors. Because women with fewer socioeconomic resources are less likely to meet recommendations for physical activity (Evenson et al., 2004), we include as predictors household income (measured as percent of the poverty line, which adjusts income for the number of household members supported on that income), years of maternal education, marital status (married or unmarried), employment status (working outside the home or not), and an indicator of black race, all of which have been associated with socioeconomic status in the PIN data. Other maternal-level predictors of interest include biologic factors (age in years and body mass index), an indicator of whether the mother is pregnant with her first child (women with previous children may be less likely to have time for leisure activities and more likely to spend time in child care activities), an indicator of diagnosis of gestational diabetes in pregnancy, and an indicator of whether the mother reported smoking cigarettes during pregnancy. Meteorological variables of interest include weekly average ozone (obtained from the US EPA, `http://www.epa.gov/ttn/airs/airsaqs/`) and weekly average temperature and precipitation (obtained from the National Climate Data Center, `http://cdo.ncdc.noaa.gov/CDO/cdo`); however, because women reported activity levels as a weekly summary, it is unclear whether meteorological variables will be as closely related to activity as they would be if daily information on activity were available.

Finally, investigators are also interested in whether activity levels were associated with physical characteristics of the neighborhood built environment, measured at the level of each woman's street segment (defined as the minimal length of street not intersected by another

7

road; e. g. the length of her block). The neighborhood environment may influence physical activity behavior, though results in the literature have been mixed (Laraia et al. 2007; Yen et al. 1998; Ross, 2000). Although neighborhood quality scales have been developed for urban neighborhoods (Caughy et al. 2001), our exploratory work suggests that these measures, developed in inner-city environments, may be inappropriate for our primarily small-town and rural study area, so that we examine the neighborhood environment through multiple individual-level indicators. These included whether the segment was urban or rural, the speed limit, whether or not the residential units were in good condition, whether or not the grounds were in good condition, whether or not abandoned residential units were present, and whether public spaces on the segment (such as parks) were in poor condition. In addition, raters of each street segment noted whether litter or graffiti were present, whether houses on the street had decorations (e. g. a wreath on the front door), whether borders (like fences or hedges) separated yards from the street, presence of signs (including neighborhood entrance signs and crime watch signs), whether most homes had porches, whether people were visible in their yards or along the street, whether there were playgrounds on the street segment, and whether a sidewalk was present along the segment. Descriptive statistics for these predictors are presented in Table 1.

3 Variable selection for the multivariate spatially-varying coefficient model

Because the observed physical activity has a high proportion of zeros (Figure 1b), we use a truncated normal model. However, exploratory analysis suggests that there is more mass at zero than can be accounted for by the truncated normal model. Section 3.1 introduces a zero-inflated truncated normal model for these data. Section 3.2 describes the spatially-varying coefficients model relating the predictors to physical activity, and Section 3.3 proposes a Bayesian variable selection algorithm to search for the subset of the predictors best fitting the data.

3.1 Zero-inflated truncated normal spatial model

Let y_{li} be the observed activity type l for subject i with spatial location (lat/long of residence) s_i, $l = 1, ..., k$ and $i = 1, ..., n$. We assume

$$y_{li} = \begin{cases} 0 & \text{with probability } p_l \\ z_{li}^+ & \text{with probability } 1 - p_l, \end{cases} \tag{1}$$

where z_{li} is a latent continuous process defined on $\mathcal{R}$, $z_{li}^+ = \max(0, z_{li})$, and $p_l \sim \text{Unif}(0,1)$ is the additional mass at zero for activity type l that can not be explained by the truncated normal model. An extension of this model would allow the probability p_l to vary with space; however, the probability of no activity under the proposed model is

$$\text{Prob}(y_{li} = 0) = p_l + (1 - p_l) * \text{Prob}(z_{li} \leq 0), \tag{2}$$

so although p_l is constant across space, $\text{Prob}(y_{li} = 0)$ is not. We assume p_l is constant across space to avoid identifiability problems.

To account for potential correlation between activity types, we model the latent process as multivariate normal

$$z_{li} = \mu_{li} + \epsilon_{li}, \tag{3}$$

where μ_{li} is the spatially-varying mean, the errors are $(\epsilon_{1i}, ..., \epsilon_{ki})' \sim \text{N}(0, \Sigma^e)$, and $\Sigma^e \sim \text{InvWishart}(k + 0.1, 0.1I)$ is the $k \times k$ cross-covariance matrix. Since the mass at zero in (2) depends on z_{li}'s density, both non-zero *and* zero observations provide information about the mean μ_{li} and covariance Σ^e.

3.2 Spatially-varying coefficients model

Spatial correlation and covariate effects are modelled in the mean using the spatially-varying coefficient model of Gelfand et al. (2003),

$$\mu_{li} = \sum_{j=0}^{p} x_{ji} \beta_{lj}(\text{s}_i) \tag{4}$$

where $x_{0i} = 1$ for all i, x_{ji} is the value of the jth covariate for subject i, and $\beta_{lj}(\text{s}_i)$ is the corresponding regression parameter for activity type l. The spatially-varying coefficient model allows the regression parameters for the jth covariate and l^{th} activity type, $\boldsymbol{\beta}_{lj} = (\beta_{lj}(\text{s}_1), .., \beta_{lj}(\text{s}_n))'$, to be different in different subregions. The regression coefficients are taken to be independent across covariates, but correlated across activity type because it is likely that covariates have a similar effect on multiple activity types. We assume $\boldsymbol{\beta}_j =$

$(\boldsymbol{\beta}'_{1j}, ..., \boldsymbol{\beta}'_{kj})'$ is a Gaussian process with mean $\boldsymbol{\theta}_j \otimes 1_n$ and separable covariance

$$Cov(\boldsymbol{\beta}_j) = K_j \otimes \Sigma^s_j,$$

where $\boldsymbol{\theta}_j = (\theta_{1j}, ..., \theta_{kj})'$, the scalar θ_{lj} is the spatial average of $\boldsymbol{\beta}_{lj}$, 1_n is the n-vector of ones, K_j is the $n \times n$ spatial correlation matrix, and Σ^s_j is the $k \times k$ covariance matrix that controls the cross-covariance between activity types.

We assume the Matern spatial correlation (Handcock and Wallis, 1994), that is, the $(i, i')^{th}$ element of K_j is

$$\frac{(2\nu_j^{1/2}||s_i - s_{i'}||/\rho_j)^{\nu_j}}{2^{\nu_j-1}\Gamma(\nu_j)}B_{\nu_j}(2\nu_j^{1/2}||s_i - s_{i'}||/\rho_j),$$

where B_{ν_j} is the modified Bessel function. The Matern correlation has two parameters: ν_j controls the smoothness (i.e., the degree of differentiability) of the process and ρ_j controls the range of spatial correlation. If $\nu_j = 0.5$ we have the exponential correlation $\exp(||s_i - s_{i'}||/\rho_j)$, and if $\nu_j = \infty$ we have the squared-exponential correlation $\exp(||s_i - s_{i'}||^2/\rho_j^2)$. We choose priors $\nu_j \sim \text{Uniform}(0.5, 2.5)$ and $\rho_j \sim \text{Uniform}(0.2, 2.5)$. Under these priors, the prior 95% interval for the spatial correlation for any pair of locations separated by 10% of the spatial domain's maximum diameter in Figure 1 is (0.01, 0.99).

3.3   Bayesian variable selection

Including all of the covariates described in Section 2 in the spatially-varying coefficient model gives an over-parameterized model. Therefore, we propose a Bayesian variable selection

model that reflects the prior belief that many of the variables should be included in the model, but only a small number have spatially-varying effects. We consider three possible relationships between activity type l and the jth covariate, x_j:

1. x_j is removed from the model and $\beta_{lj}(s_i) = 0$ for all s_i,

2. x_j's effect is constant across space and $\beta_{lj}(s_i) = \theta_{lj}$ for all s_i, or

3. x_j's effect varies spatially and $\beta_{lj}(s_i)$ varies from location-to-location.

Our Bayesian variable selection routine uses stochastic search variable selection to compute the posterior probability of each of these three relationships.

The overall average effect θ_{lj} is given the spike and slab prior, popularized for variable selection in the usual linear regression setting by George and McCullough (1993). That is,

$$\theta_{lj} = \gamma_{1lj}\alpha_{lj}, \tag{5}$$

where $\gamma_{1lj} \in \{0, 1\}$ and $\alpha_{lj} \sim N(0, \sigma_j)$. If $\gamma_{1lj} = 0$ then $\theta_{lj} = 0$ and x_j's average (over space) effect on activity type l is zero.

To determine which regression parameters should be allowed to vary spatially, the prior for the columns and rows of the covariance matrix Σ_j^s also have positive mass at zero. Let

$$\Sigma_j^s = \text{diag}(\gamma_{21j}, ..., \gamma_{2kj})\Omega_j \text{diag}(\gamma_{21j}, ..., \gamma_{2kj}), \tag{6}$$

where $\gamma_{2lj} \in \{0, 1\}$ and $\Omega_{lj} \sim \text{InvWishart}(k + 0.1, 0.1I_k)$. If $\gamma_{2lj} = 0$ then $\boldsymbol{\beta}_{lj} = \theta_{lj}$ for all s_i and the effect is x_j is constant across space for activity type l.

Since we believe there is always some spatial correlation in the residual activity due to missing covariates, we fix γ_{1l0} and γ_{2l0} to be one for all l. The prior for remaining binary variable inclusion indicators ensures that the effects are not allowed to vary spatially unless the overall average is non-zero. We choose the prior

$$
p(\gamma_{1lj}, \gamma_{2lj}) = \begin{cases} 1 - \pi_{1j}, & \gamma_{1lj} = 0 \text{ and } \gamma_{2lj} = 0 \\ \pi_{1j}(1 - \pi_{2j}), & \gamma_{1lj} = 1 \text{ and } \gamma_{2lj} = 0 \\ \pi_{1j}\pi_{2j}, & \gamma_{1lj} = 1 \text{ and } \gamma_{2lj} = 1 \end{cases} \tag{7}
$$

In this prior, π_{1j} is the prior probability that x_j is included in the model (either constant or spatially-varying) for activity type l and π_{2j} is the conditional probability that the effect of x_j varies spatially given that it is included in the model for activity type l.

Note that the prior probabilities π_{1j} and π_{2j} and variance σ_j^2 are shared across activity types. The hyperpriors are $\pi_{1j} \sim \text{U}(0,1)$, $\pi_{2j} \sim \text{U}(0,1)$, and $\sigma_j^2 \sim \text{InvGamma}(0.5,0.5)$; this hierarchical model facilitates pooling information regarding the importance of x_j across activity types and encourages variable to be selected consistently across activity types.

4   Computing methods

To facilitate MCMC sampling, we introduce auxiliary variables g_{li} which indicate whether the y_{li} is drawn from the point mass of zero or the truncated normal model in (1), that is, $g_{li} \sim \text{Bern}(p_l)$ and

$$
y_{li}|g_{li} = \begin{cases} 0 & \text{if } g_{li} = 1 \\ z_{li}^+ & \text{if } g_{li} = 0. \end{cases} \tag{8}
$$

g_{li} is updated via Gibbs sampling; for observations with $y_{li} > 0$ g_{li} is fixed at zero, for observations with $y_{li} = 0$ the full conditional probability that $g_{li} = 1$ is $p_l / (p_l + (1 - p_l) * Prob(z_{li} < 0))$. The proportion p_l is updated from its full conditional $p_l \sim \text{Beta}(1 + \sum g_{ji}, 1 + n - \sum g_{ji})$.

The latent variables z_{li} for observations with $y_{li} = g_{li} = 0$ are drawn using Gibbs sampling from their truncated normal full-conditionals; if $g_{li} = 1$ z_{li} no longer appears in the likelihood and it is updated from its normal prior. The covariance matrices Σ^e and Ω_{lj} are drawn using Gibbs sampling from their inverse Wishart full-conditionals and ρ_j and ν_j are drawn using Metropolis-Hastings sampling with Gaussian candidate distributions. Candidates with no prior mass are simply rejected.

For computational convenience, we rewrite the spatially-varying mean as

$$\mu_{li} = \sum_{j=0}^{p} x_{ji} \left[\gamma_{1lj} \alpha_{lj} + \gamma_{2lj} \beta_{lj}(s_i) \right], \tag{9}$$

where $\alpha_{lj} \sim \text{N}(0, \sigma_\alpha^2)$, $\boldsymbol{\beta}_j \sim \text{N}(0, K_j \otimes \Omega_j)$. The parameters in μ_{li} are updated using blocked-Gibbs sampling with blocks $\Theta_{lj} = \{ \gamma_{1lj}, \gamma_{2lj}, \alpha_{lj}, \boldsymbol{\beta}_{lj} \}$. To derive the full-conditionals, let $\Theta_{(lj)}$ be the set of all parameters not included in Θ_{lj}. We draw from $p(\Theta_{lj} | z, \Theta_{(lj)})$ by first integrating over $(\alpha_{lj}, \boldsymbol{\beta}_{lj})$ and drawing $(\gamma_{1lj}, \gamma_{2lj})$ from $p(\gamma_{1lj}, \gamma_{2lj} | z, \Theta_{(lj)})$ and then drawing $(\alpha_{lj}, \boldsymbol{\beta}_{lj})$ from $p(\alpha_{lj}, \boldsymbol{\beta}_{lj} | \gamma_{1lj}, \gamma_{2lj}, z, \Theta_{(lj)})$.

Define the residuals (not removing the x_j's effect on the lth activity type) as $r_{li} = z_{li} - \mu_{li} + x_{ji} \left[\gamma_{1lj} \alpha_{lj} + \gamma_{2lj} \beta_{lj}(s_i) \right]$ and simply $r_{l'i} = z_{l'i} - \mu_{l'i}$ for $l' \neq l$. Also, let $U^{-1} = \Sigma^e$ and $V^{-1} = \Omega_j$. Routine algebra shows that

$$p(\gamma_{1lj}, \gamma_{2lj} | z, \Theta_{(lj)}) \propto |\Sigma_r|^{1/2} \exp \left(\frac{1}{2} \hat{\boldsymbol{\beta}}' \Sigma_r \hat{\boldsymbol{\beta}} \right) p(\boldsymbol{\gamma}_{1j}, \boldsymbol{\gamma}_{2j}), \tag{10}$$

14

where $\Sigma_r = (U_{ll}\Gamma X'X\Gamma + \Delta^{-1})^{-1}$,

$$\Delta = \begin{pmatrix} \sigma_\alpha^2 & 0 \\ & \\ 0 & V_{ll}K_j \end{pmatrix}, \quad X = \begin{pmatrix} x_{j1} & x_{j1} & & 0 \\ \vdots & & \ddots & \\ x_{jn} & 0 & & x_{jn}) \end{pmatrix},$$

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 0 \\ \\ \Gamma X' \sum_{l'=1}^{k} U_{ll'}\mathrm{r}'_{l'} + \Omega \sum_{l' \neq l}^{k} V_{ll'}\boldsymbol{\beta}'_{l'j} \end{pmatrix}$$

and Γ is the $(n+1) \times (n+1)$ diagonal matrix with diagonal elements $\gamma_{1lj}, \gamma_{2lj}, ..., \gamma_{2lj}$. The posterior probability of each $(\delta_{1lj}, \delta_{2lj})$ combination is computed and $(\delta_{1lj}, \delta_{2lj})$ is drawn accordingly. Given $(\delta_{1lj}, \delta_{2lj})$, $(\alpha_{lj}, \boldsymbol{\beta}'_{lj})'$ is drawn from a multivariate normal distribution with mean $\Sigma_r\hat{\boldsymbol{\beta}}$ and covariance Σ_r.

When fitting this model and simplifications of it, we compare models using the deviance information criterion (DIC) of Speigelhalter et al. (2002), defined as $DIC = \bar{D} + p_D$ where $\bar{D}$ is the posterior mean of the deviance, $p_D = \bar{D} - \hat{D}$ is the effective number of parameters, and $\hat{D}$ is the deviance evaluated at the the posterior mean of the parameters in the likelihood. The deviance is calculated using the marginal likelihood of y_{li} given μ_{li}, Σ^e, and p_l after analytically integrating over the latent variables z_{li} and g_{li}. The model's fit is measured by $\bar{D}$, while the model's complexity is captured by p_D. Models with smaller DIC are preferred.

5   Analysis of the PIN3 data

We begin our analysis of the PIN3 data by comparing the full model first described in Section 3 (Model 1) with several simpler alternatives. This is a multivariate model for activity in

the six domains (work, recreational, outdoor household, indoor household, adult and child care, and transportation) as a function of demographic and other person-level variables as well as meterological and neighborhood-level variables.

Table 2 shows that the full model has a smaller DIC than the models without spatially-varying effects (Model 2), correlation across activity types (Model 3), the variable selection hierarchy (Model 4), and the model without all three of these features (Model 5). Model 2 has the smallest DIC of the simplified models. Eliminating the possibility of spatially-varying effects gives a smaller model (p_D=97.7 compared to $p_D = 153.6$ for the full model) but less fidelity to the data ($\bar{D}$=13451.3 compared to $\bar{D} = 13392.8$ for the full model). Balancing these two factors gives a slightly smaller DIC for the full model (DIC=13546.4) compared to the simplified Model 2 (DIC=13549.0). As discussed further below, only a small number of variables have significant probability of varying spatially. Including the unnecessary variables inflates Model 1's DIC. DIC reduces to 13536.0 when the full model is refit allowing only the subset of variables! identified by Model 1 as spatially-varying (probability greater than 0.5) to vary spatially.

The model without the variable selection hierarchy (i.e., Model 4 with $\pi_{1j} = \pi_{2j} = 0.5$ and $\sigma_j = 1$ for all j) also has DIC similar to the full model. To illustrate the difference between these models we plot the overall (spatially-varying and constant) inclusion probabilities for each activity type and covariate in Figure 2. The full model's inclusion probabilities are more consistent across activity types. For example, employment status is included for 4 of the 6 activity types under Model 4. All six inclusion probabilities are higher under the full model and employment status is now included for all 6 activity types. Also, borrowing strength across activity types removes some variables with inclusion probabilities just over

0.5 for one activity type (precipitation, bad residential units, and bad public places), which are likely to be spurious associations.

Table 3 gives the 95% intervals for the zero-inflation parameters and the residual correlations. Outdoor household and transportation-related activity have the largest proportions of zero in the data set (84% for outdoor household and 78% for transportation), and these activities generally have the largest zero-inflation parameters p_l (95% intervals (0.29, 0.42) for outdoor household and (0.16, 0.29) for transportation) which represent the excess mass at zero not explained by the truncated normal model. For the other four types of activity, a truncated Gaussian distribution fits fairly well, and the zero-inflation parameters are small. The residual correlation between indoor household activity and all other activities except transportation activity are significantly positive, and no other correlations are significantly non-zero. Therefore indoor household activity may serve as the best indicator of women's overall activity level.

As expected, individual-level factors tended to be more strongly associated with physical activity levels than neighborhood environmental factors, with several factors having posterior inclusion probabilities > 0.90. Education, age, number of previous births, and employment status are important predictors of all six activity types (Figure 2a). In particular, more highly educated women tend to have higher levels of recreational, outdoor household, care, and transportation activity and lower levels of work-related and indoor household activity. These women may be more likely to have sedentary jobs and to hire domestic workers to help with household chores, freeing more time for leisure activity and playing with children. Older women are more likely to have greater activity levels in every domain except for work activity, which they are less likely to report than are their younger counterparts. Women

17

who have previously given birth are more likely to report child and adult care activity as well as indoor and outdoor household activity; they are less likely to report transportation, recreational, and work-related activity. Women working outside the home are more likely to report transportation, outdoor household, and work activity and are less likely to report activity in the other domains. The data show that activity patterns differ by race; black women generally exert more energy on work and transportation-related activities and less energy on recreational, outdoor household, and child and adult care activities. Also, women with high BMI are less likely to participate in recreational or outdoor household activities. Gestational diabetes and smoking were associated only with increased work activity, which is perhaps a reflection of the demographics of these conditions. Women with higher incomes tended to have more recreational and less transportation activity. Marital status was unassociated with activity levels in any domain.

Ozone is a predictor of recreational, outdoor household, and child and adult care activities. The posterior mean of the ozone coefficient is positive for recreational and outdoor household activity, that is, women are more likely to be active outdoors when the ozone is high. This is somewhat surprising given the well-documented health effects of ozone exposure (National Research Council, 2008), especially for susceptible groups. To investigate this issue further, we refit the model with separate ozone effects for women with and without at least 16 years of education. For the high education group ozone is only included with probability higher than 0.5 for child and adult care (probability 0.51) and the posterior mean for the overall average effect α for child and adult care is -0.10, which corresponds to roughly a 10% decrease in MET minutes per week ($\exp(-.1)=.9$) for a standard deviation (1.04 ppm) unit increase in ozone. In contrast, for the low education group ozone is an important predictor

18

for only outdoor household activity (probability 0.71) and the posterior mean for the overall average effect is 0.16, which corresponds to roughly a 17% increase in MET minutes per week for a standard deviation unit increase in ozone. Therefore, the association between physical activity and ozone is dependent on education level. In addition, ozone levels tend to be higher in summer, when recreational and outdoor household activity is more frequent, and given that reporting is only on a weekly basis, presence of high ozone for the week could very well be a marker of a week with nice weather and only one red or orange ozone day. Along these lines, higher temperatures were associated with more recreational and outdoor household activity, and lower child and adult care activity (perhaps because children participated in recreational activity).

Several neighborhood characteristics are also identified as important predictors of activity. Women in neighborhoods with abandoned residential units engage in more indoor, child care, and transportation activity, perhaps a sign of lower socioeconomic status. Also, outdoor household activity increases in neighborhoods with porches and no sidewalks. This may be because such lots are more common in rural areas in which yards are larger and opportunities for outdoor household activities, such as gardening, are more plentiful. However, the presence of a sidewalk increases the average amount of transportation activity, perhaps because mothers are more likely to walk or bike to work or other places.

Figure 3 plots the posterior probability that each covariate is a spatially-varying effect. As expected, only a small number of the covariates have spatially-varying effects. Two associations, employment status with working activity and the first birth indicator with adult and child care activity, have inclusion probability greater than 0.9. This indicates a strong signal from the data because the prior probability of a spatially-varying effect is

only 0.25. Figure 4 shows that working women are generally more activity in working and transportation throughout the spatial domain. The association between employment status and work-related activity (Figure 4a) is strongest in urban areas and weakest in rural areas, perhaps because women in rural areas may consider some outdoor household activities on farms to be work-related activities. Employment is also identified as having a spatially-varying effect on transportation activity, with a stronger effect (Figure 4b) in the highly bikeable areas of Alamance and western Orange counties. While not having given birth previously was strongly protective against child and adult care activity, this effect (Figure 4c) was weaker in rural areas of northern Chatham and southern Alamance counties, in which women may be more likely to care for other family members.

Several other variables have posterior probabilities greater than 0.5 of having a spatially-varying effect on activity. For example, Figure 4d shows the negative association between BMI and recreational activity is the strongest in the Chapel Hill area (southeast Orange County) and in Snow Camp (southwest Alamance county) compared to northern Orange or Durham counties. This could be due to unique cultural factors related to life in the college town of Chapel Hill or to the Quaker religious community in Snow Camp.

None of the neighborhood variables are included as spatially-varying effects with probability greater than 0.5. The variable most likely to be included is the effect of porch on outdoor activity. This variable is included in the model with probability 0.67 and is included as a spatially-varying effect with probability 0.35. Therefore the conditional probability of being spatially-varying given inclusion in the model is slightly higher than 0.5 (Figure 3b). Figure 5 shows that the positive effect of having a porch is the largest in rural Chatham County where a porch may indicate a house with a large property that requires upkeep.

6 Discussion

In this paper we developed a Bayesian hierarchical model to analyze data from an epidemiologic study designed to assess whether neighborhood environmental characteristics had a noticeable effect on physical activity levels of pregnant women. To adequately model these complex data, we develop a spatial model for semicontinuous data which allows the effect of a subset of the predictors to vary over the spatial domain. We found that while some neighborhood characteristics were associated with physical activity levels, these associations were not particularly strong, especially in relation to individual characteristics of the women under study, such as age or level of education.

Another important direction of research in this area is to identify factors that lead to a change in activity from pre-pregnancy to postpartum. We cannot fully address this question with these data because we do not have pre-pregnancy activity data. However, the survey does include a self-reported binary indicator of whether the mother regularly exercised before pregnancy. When we included this variable as a predictor in our hierarchical model it was an important predictor of recreational activity only, and the associations between the other predictors and the outcome remained largely the same.

Other future work involves examining determinants of change in physical activity levels from pregnancy to the postpartum. In this study, a subset of subjects provided physical activity data when their children were 3 and 12 months of age. While neighborhood environmental characteristics were not strongly associated with activity levels during pregnancy, we anticipate some different trends in the postpartum period. For example, a woman might be unlikely to walk or jog with her infant in a stroller without having a sidewalk. We are

21

interested in extending the methods to account for longitudinal, multivariate responses in order to determine whether neighborhood environmental characteristics may be related to change (or retention) of activity levels during the postpartum period.

References

Ainsworth, B., Haskell, W., Whitt, M., Irwin, M., Swartz, A., Strath, S., O'Brien, W.L., Bassett, D.R., Schmitz, K.H., Emplaincourt, P.O., Jacobs, D.R., and Leon, A.S. (2000). Compendium of physical activities: an update of activity codes and MET intensities. *Medicine and Science in Sports and Exercise* 32(9 Suppl.):S498-S516.

Albert PS, Shen J (2005). Modelling longitudinal semicontinuous emesis volume data with serial correlation in an acupuncture clinical trial. *Applied Statistics*, **54**, 707–720.

Agarwal DK, Gelfand AE, Citron-Pousty S (2002). Zero-inflated models with application to spatial count data. *Environmental and ecological statistics*, **9**, 341.

Berk KN, Lachenbruch PA (2002). Repeated measures with zeros. *Statistical Methods in Medical Research*, **11**, 303–316.

Caughy MO, O'Campo PJ, Patterson, J (2001). A brief observational measure for urban neighborhoods. *Health and Place*, **7**, 225-236.

Chipman H (1996). Bayesian variable selection and related predictors. *Canadian Journal of Statistics*, **24**, 17–36.

Clapp J III, Little K (1995). Effect of recreational exercise on pregnancy weight gain and subcutaneous fat deposition. *Medicine and Science in Sports and Exercise*, **27**, 170-177.

Dempsey JC, Sorensen TK, Williams MA, Lee I-M, Miller RS, Dashow EE, Luthy DA (2004). Prospective study of gestational diabetes mellitus risk in relation to maternal recreational physical activity before and during pregnancy. *American Journal of Epidemiology*, **159**, 663-670.

Evenson, K.R., Savitz, D.A., and Huston, S.L. (2004). Leisure-time physical activity among pregnant women in the U.S. *Paediatric and Perinatal Epidemiology* 18(6):400-407.

George EI, McCulloch RE (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

George EI, McCulloch RE (1997). Approaches for Bayesian variable selection, *Statistica Sinica*, **7**, 339–373.

Gelfand AE, Kim HK, Sirmans CF, Banerjee S (2003). Spatial modelling with spatially varying coefficient processes. *Journal of the American Statistical Association*, **98**, 387–396.

Handcock MS, Wallis JR (1994). An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). *Journal of the American Statistical Association*, **89**, 368–390.

Hoeting JA, Davis RA, Merton AA, Thomspon SE (2006). Model Selection for Geostatistical Models. *Ecological Applications*, **16**, 87–98.

Huang HC, Chen CS (2007). Optimal Geostatistical Model Selection. *Journal of the American Statistical Association*, **102**, 1009-1024.

Kelsall JE, Diggle PJ (1998). Spatial variation in risk of disease: a nonparametric binary regression approach. *Journal of the Royal Statistical Society, Series C*, **47**, 559-573.

Laraia B, Messer L, Evenson K, and Kaufman JS (2007). Neighborhood factors associated with physical activity and adequacy of weight gain during pregnancy. *Journal of Urban Health*, **84**, 793-806.

Mitchell TJ, Beauchamp JJ (1988). Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association*, **83**, 1023–1036.

National Research Council, Committee on Estimating Mortality Risk Reduction Benefits from Decreasing Tropospheric Ozone Exposure (2008). Estimating Mortality Risk Reduction and Economic Benefits from Controlling Ozone Air Pollution. *National Academy of Sciences*.

Olsen MK, Schafer JL (2001). A two-part random effect model for semicontinuous longitudinal data. *Journal of the American Statistical Society*, **96**, 730–745.

Paciorek CP (2007). Computational techniques for spatial logistic regression with large data sets. *Computational Statistics & Data Analysis*, **51**, 3631–3653.

Reich BJ, Hodges JS, Zadnik V (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, **62**, 1197–1206.

Ross C (2000). Walking, exercising, and smoking: does neighborhood matter? *Social Science and Medicine*, **51**, 265-274.

Saftlas AF, Logsden-Sackett N, Wang W, Woolson R, Bracken MB. (2004). Work, leisure-time physical activity, and risk of preeclampsia and gestational hypertension. *American Journal of Epidemiology*, **160**, 758-765.

Smith M, Fahrmeir, L (2007). Spatial Bayesian Variable Selection With Application to Functional Magnetic Resonance Imaging. *Journal of the American Statistical Association*, **102**, 417-431.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002). Bayesian measures of model complexity and fit (with discussion and rejoinder). *J. Roy. Statist. Soc., Ser. B*, **64**, 583–639.

Stein ML (1992). Prediction and Inference for Truncated Spatial Data. *Journal of Computational and Graphical Statistics*. **1**, 91-110.

Yen IH, Kaplan GA (1988). Poverty area residence and changes in physical activity level: evidence from the Alameda County Study. *American Journal of Public Health*, **88**, 1709-1712.

Zhang M, Strawderman RL, Cowen ME, Wells MT (2006). Bayesian inference for a two-part hierarchical model: An application to profiling providers in managed health care. *Journal of the American Statistical Society*, **101**, 934–945.

Table 1: Summary statistics for the predictor variables in PIN3 data. Only the sample mean (proportion) is given for binary variables.

Variable	Mean	SD	Variable	Mean	SD
Percent of poverty line	439.43	218.58	Bad residential units	0.04	–
Maternal education	16.17	2.76	Bad residential kept grounds	0.10	–
Black	0.15	–	Abandoned residential units	0.03	–
Married	0.82	–	Bad public spaces	0.07	–
Age (years)	29.71	4.59	Litter	0.44	–
First birth	0.49	–	Graffiti	0.02	–
BMI	25.13	6.38	Decoration	0.78	–
Gestational diabetes	0.04	–	Border	0.26	–
Smoker	0.09	–	Entrance sign	0.12	–
Working	0.71	–	Crime watch	0.08	–
Ozone (ppm)	2.85	1.08	Porches	0.78	–
Temperature (F)	60.87	15.28	People	0.39	–
Precipitation (In)	0.12	0.13	Playground	0.05	–
Urban	0.70	–	Church playground	0.01	–
Speed limit (MPH)	28.69	8.41	Sidewalk	0.35	–

Table 2: Model comparison via DIC. Model 1 is the full model described in Section 3. Models 2-4 simplify the Model 1 in various ways. Model 2 does not allow spatially-varying coefficients (i.e., $\gamma_{2j} = 0$), Model 3 removes correlation across activity type (i.e., Σ^e and Ω are diagonal), Model 4 removes the variable selection hierarchy (i.e., $\pi_{1j} = \pi_{2j} = 0.5$ and $\sigma_j = 1$), and Model 5 removes all three of these features.

Model	$\bar{D}$	p_D	DIC
1	13392.8	153.6	13546.4
2	13451.3	97.7	13549.0
3	13449.5	128.2	13577.6
4	13348.6	203.0	13551.6
5	13497.7	98.9	13596.5

Table 3: Summary of the zero-inflation parameters and residual correlation. Table (a) gives the sample proportion of zeros the posterior 95% intervals of the zero-inflation parameters p_l. Table (b) gives the 95% intervals for the elements of $D^{-1/2}\Sigma^e D^{-1/2}$, where D is $k \times k$ diagonal matrix with elements Σ_{ll}^e.

(a) Zero-inflation parameters p_l

	Working	Rec	Outdoor	Indoor	Care	Trans
Sample	0.69	0.37	0.84	0.34	0.62	0.78
p_l	(0.00, 0.11)	(0.00, 0.01)	(0.29, 0.42)	(0.00, 0.01)	(0.02, 0.15)	(0.16, 0.29)

(b) Residual correlation parameters in Σ^e

	Rec	Outdoor	Indoor	Care	Trans
Work	(-0.09, 0.09)	(-0.08, 0.18)	(0.07, 0.24)	(-0.11, 0.24)	(-0.06, 0.17)
Recreational		(-0.02, 0.19)	(0.00, 0.14)	(-0.01, 0.14)	(-0.05, 0.14)
Outdoor			(0.04, 0.24)	(-0.05, 0.24)	(-0.15, 0.10)
Indoor				(0.06, 0.24)	(-0.14, 0.04)
Care					(-0.06, 0.16)

Figure 1: Plots of the PIN data. Panel (a) gives the spatial locations (plus "jitter") of the pregnant women's residences and Panel (b) plots log(transportation-related activity +1).

(a) Spatial locations
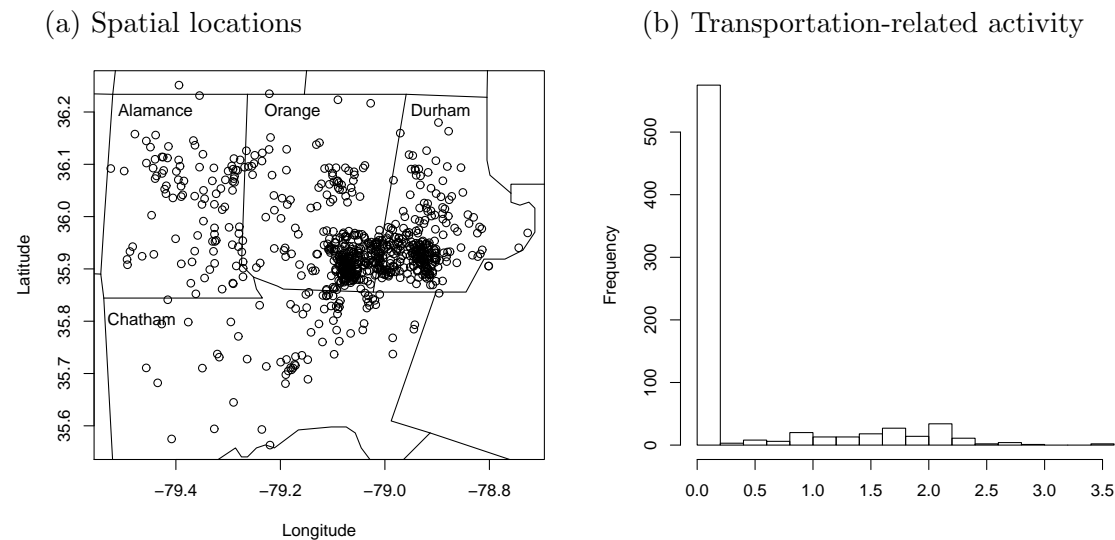
(b) Transportation-related activity

25

Figure 2: Posterior inclusion probabilities (mean of γ_{1jl}) for the full model (Panel (a)) and the model without the variable selection hierarchy (i.e., with $\pi_{1j} = \pi_{2j} = 0.5$ and $\sigma_j = 1$, Panel (b)). Gray and block boxes represent probabilities greater than 0.5 and 0.9, respectively. The "+" and "-" for the inclusion probabilities indicate the sign of the posterior mean of overall average effect θ_{jl}.
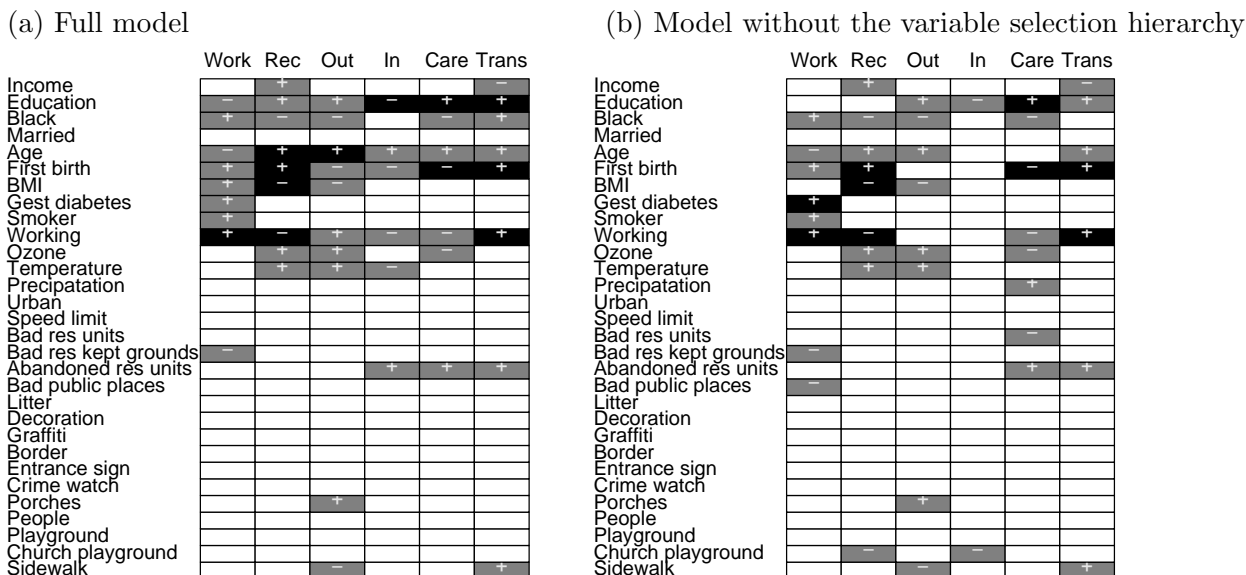

(a) Full model

(b) Model without the variable selection hierarchy

Figure 3: Posterior probabilities of being selected as a spatially-varying coefficient (mean of γ_{2jl}) and the conditional probability of being included with a spatially-varying coefficient given the variable is included in the model (mean of γ_{2jl}/mean of γ_{1jl}). Gray and block boxes represent probabilities greater than 0.5 and 0.9, respectively.
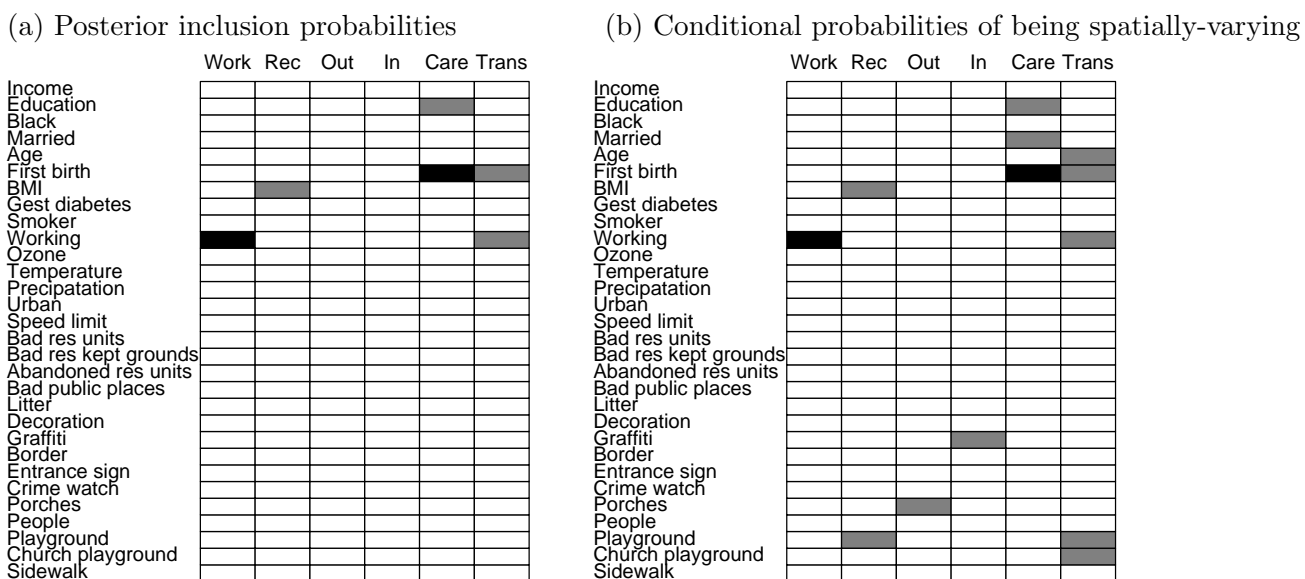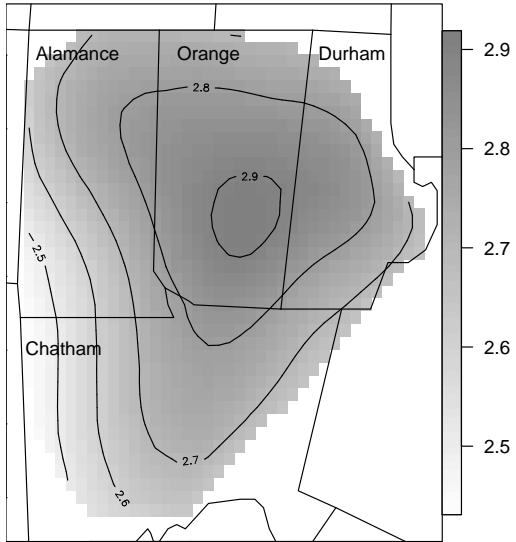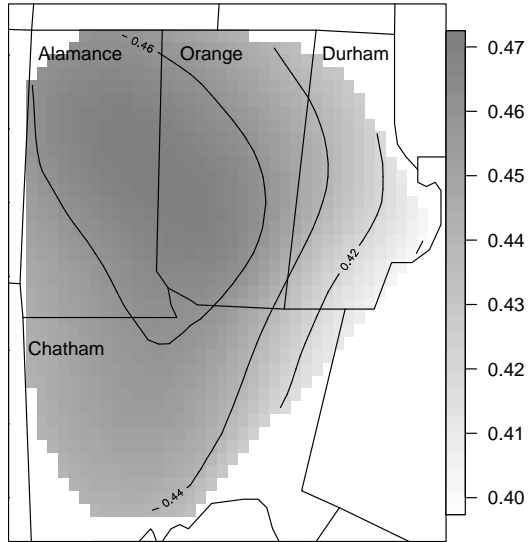

(a) Posterior inclusion probabilities

(b) Conditional probabilities of being spatially-varying

Figure 4: Posterior mean of the spatially-varying effects ($\boldsymbol{\beta}$)
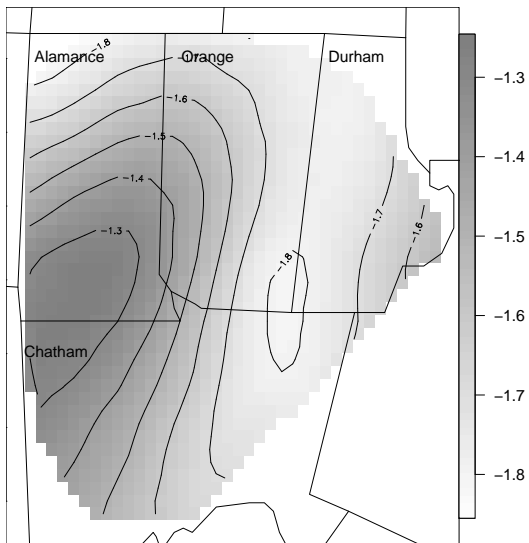
(a) Employment status on work activity

(b) Employment status on transportation activity

(c) First birth on care activity
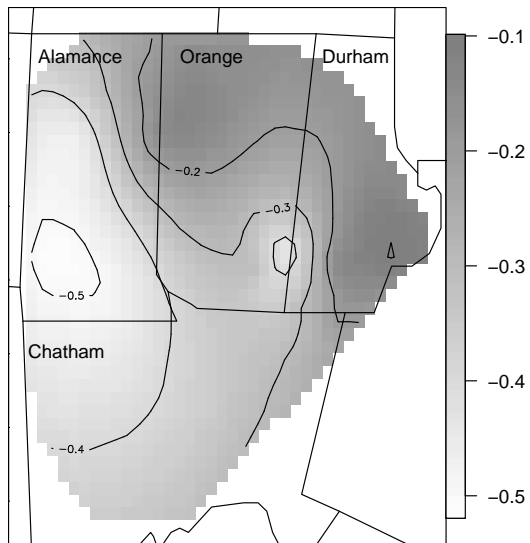
(d) BMI on recreational activity

Figure 5: Posterior mean of the spatially-varying effect ($\boldsymbol{\beta}$) of having a porch on outdoor activity