Continuous and Discrete Spatial Heterogeneity: Modeling Strategies and Simulation.

Mauricio Sarrias*

May 25, 2015

Abstract

Continuous and discrete unobserved heterogeneity have been widely used in modeling discrete choice models. This paper shows how these modeling strategies can be used to capture and model spatial heterogeneity or locally varying coefficients for different latent structures. We also outline the main advantages and disadvantages of both models. Then, we conduct a simulation experiment in order to understand the ability of both approaches to retrieve the true representation of the spatially varying process under small sample size situations. Our results show that the data requirement to achieve lower bias in the continuous case is substantial compared with the discrete case. We also found that, as the number of individuals per spatial unit increases, both models are able to identify the regional-specific estimates. However, the discrete case is able to retrieve the true spatial heterogeneity surface with lower bias and better coverage.

JEL classification: C13, C15, C21, C25
Keywords: Spatial Heterogeneity, Simulation, Random Parameters, Spatial Econometrics

*Department of City and Regional Panning, Cornell University, Ithaca, NY, USA, 14853; Email: ms2822@cornell.edu

1 Introduction and Background

In the last decades has been a growing interest in local analysis where the aim is to model spatial heterogeneity in the form of spatially varying coefficients (see for example Fotheringham and Brunsdon, 1999; Fotheringham, 1997; Lloyd, 2010, for a review). Anselin (1988, page 9) defines spatial heterogeneity in econometric terms as the *"structural instability over space, in the form of different response functions or systematically varying parameters"*. Formally, we might define spatial heterogeneity as $y_c = f_c(x_c) + \epsilon_c$, $c = 1, ..., C$ where c is the index for spatial location; y_c is the dependent variable, x_c is the predictor variable measured at location c; ϵ_c is the error term, and $f_c(x_c)$ indicates that the structural relationship between y and x varies across spatial units. If we are willing to assume that this relationship is linear, then we might write $f_c(x_c) = \beta_c x_c$. Thus, the response to a particular variable, β_c, is inherently different across space and thus spatial nonstationary (Fotheringham et al., 1996; Brunsdon et al., 1998a,b).

If spatial heterogeneity is present in the data[1], then global models that assume that the relationship between y_c and x_c is constant across geographical space may not be entirely appropriate. For example, a frequently used procedure is to regress y_c on x_c using the full spatial data set. Hence, $\widehat{\beta}$ give us the 'average' correlation between both variables for all locations: if the coefficient is positive (negative) we would say that an increase (decrease) in x_c is correlated with an increase in y_c regardless of the geography location, c. As Ali et al. (2007) point out, if the goal of the researcher is to test average effects (global parameters), or to provide benchmarks to establish stylized fact, then global models and standard approaches are quite suitable. Nevertheless, these approaches may obscure significant spatial variation and hide important local differences resulting in misleading and inadequate policy inferences for spatial units that follow a completely different process than the average pattern (Brunsdon et al., 1998a; Fotheringham, 1997; Charlton and Brunsdon, 1997).

Several statistical approaches have been developed to account for spatially varying relationships. These methods enable estimation of model parameters locally, or they allow model parameter to vary as a function of location.[2] One of them is the spatial expansion method (SEM) (Casetti, 1972; Kochanowski, 1990; Casetti, 1997; Casetti and Jones III,

[1]There are several reasons why we might expect relationships to vary over space (see for example Fotheringham et al., 2003, chap. 1 for a more complete discussion). The fist variation, and less interesting, is sampling variation. This occurs when spatial variation is because of using different sample data. The second reason is that some relationships are intrinsically different across space. For example, there are spatial variations in people's attitudes or preferences or there are different contextual issues that produce different responses to the same stimuli over space. The third possible cause of spatial heterogeneity is omitted variables or incorrect functional form

[2]For further review see for example Fotheringham and Brunsdon (1999), chapter 6 in Anselin (1988), Lloyd (2010), and chapter 1 in Fotheringham et al. (2003).

2003; Brown and Jones, 1985; McMillen, 1996). This method allows the coefficients to be a function of geographical coordinates in the following way:

$$\begin{aligned} y_c &= \alpha_c + \beta_c x_c + \epsilon_c, \quad c = 1, ..., C \\ \alpha_c &= \alpha_0 + \alpha_1 u_c + \alpha_2 v_c \\ \beta_c &= \beta_0 + \beta_1 u_c + \beta_2 v_c \end{aligned}$$

where u_c and v_c represent the spatial coordinates of location c. Specifications of the parameters represent simple linear expansions of the global parameter over space but more complex expansion as nonlinear and quadratic expression can be accommodated (see for example Fotheringham et al., 2003; Páez, 2005). This model can be easily estimated by OLS or nonlinear regression depending on the nature of y and the distribution of the error term. The main disadvantage of this method is that the form of the expansion needs to be assumed a priori and also in a deterministic way (Fotheringham et al., 2003). To address the latter problem, one can assume that the local relationship varies randomly over geographical space (Swamy, 1971). This method is known as the random coefficient model (RCM). In this method, the coefficients are assumed to be normally distributed, i.e., $\beta_c \sim \mathrm{N}(0, \sigma_\beta^2)$ where σ_β is estimated.[3]

The geographically weighted regression (GWR) (Brunsdon et al., 1998b; Fotheringham et al., 2003, 2009) represents one of the most promising approaches to address spatial nonstationarity. Unlike the previous two methods, GWR approach computes local relationships that vary smoothly over space by considering how they behave in the vicinity, and hence taking into account spatial location explicitly. Coefficients are not assumed to be random, but a function of the coordinates in geographical space of the cth observation:

$$y_c = \mathbf{x}_c' \boldsymbol{\beta}(u_c, v_u) + \epsilon_c, \quad c = 1, ..., C$$

The principle of GWR is to place a kernel around location c, and estimate the local coefficients using the information located inside the kernel window. These parameters are intended to reflect the spatial heterogeneity in the sample by estimating different marginal responses to an explanatory variable across space. The estimation is carried out by geographically iteratively weighted least squares, with weights based on any kernel distribution as function of the distance between region c and the nearby points. Thus, observations closer to c would have more weights and greater impact on parameter estimates than those farther away.

The three methods presented above share an important limitation. These approaches require aggregating the variables at the location level. Therefore, we are prevented

[3] A nonparametric extension of the technique is to drop the normality assumption on the coefficient and to estimate the distribution itself from the data (see Aitkin, 1996, for further details).

from using data at the individual level. This raises concerns about the misleading conclusions that can be derived at the individual level by using aggregate variables known as the 'ecological fallacy problem' (Robinson, 1950; Peeters and Chasco, 2006; Anselin, 2002).

A widely used methodology that avoids this problem by combining both individual-level and aggregate (contextual) variables is multilevel modeling.[4] This methodological approach permits to separate the effect of personal and place characteristics on behavior and to investigate the extent and nature of spatial variation in individual outcomes measures (Goldstein, 1987; Jones, 1991; Duncan and Jones, 2000). The structural model is:

$$\begin{aligned} y_{ic} &= \alpha_c + \beta_c x_{ic} + \epsilon_{ic}, \quad i = 1, ..., N;\ c = 1, ..., C \\ \alpha_c &= \alpha_0 + u_{\alpha,c} \\ \beta_c &= \beta_0 + u_{\beta,c} \end{aligned}$$

where $u_{\alpha,c} \sim \mathrm{N}(0, \sigma^2_\alpha), u_{\beta,c} \sim \mathrm{N}(0, \sigma^2_\beta)$ and $\epsilon_{ic} \sim \mathrm{N}(0, \sigma^2_\epsilon)$. Like the RCM, the parameters vary randomly across regions. Several improvements to the multilevel modeling have been made. For example, including place attributes in the constant and the coefficients, and extending the number of nested levels in the hierarchy beyond two.

The main drawback of multilevel modeling is that usually the random coefficients are assumed to be normal distributed. This makes easier the estimation process, but creates other problems. The assumption of a normal distribution implies that some locations might have positive or negative coefficients, whether or not this is true.[5] In practice, this implies that occasionally researchers find sign reversals that are counterintuitive and difficult to explain. Furthermore, the domain of the normal distribution is $(-\infty, +\infty)$, which results in unreliable extreme coefficients and high coefficient variability. Those problems have been also found when applying the GWR approach (See for example Jetz et al., 2005; Páez, 2005; Páez et al., 2011).

Under this context, this study has two purposes. First, we expand the tool kit for modeling spatially varying coefficients by introducing discrete and continuous spatial heterogeneity. Both modeling strategies are intended to complement the existing approaches by using variables at the individual level, and can be applied to a vast array of behavioral-spatial questions. We mainly borrow some modeling ideas that have been widely used in discrete choice modeling, concretely in the multinomial logit context[6], and by doing so, we assume

[4] For other quantitative methods that avoid the ecological fallacy problem see Withers (2001).

[5] See Section 2 for more discussion about the normal assumption.

[6] For a further review on the multinomial logit with random parameters, also known as the mixed logit model, see Train (2009) and Hensher and Greene (2003).

that the parameters $\boldsymbol{\beta}_c$ vary randomly across spatial units according to some distribution $g()$. The two methods differ on how the underlying distribution $g()$ is approximated. The approximation can be summarized as follows: 1) If we assume that the spatial heterogeneity is continuous, that is, the coefficients in each location can take any real number in some interval, then $g()$ can be any continuous distribution. The choice of $g()$ will mainly depend on the assumptions about the domain and boundedness of the coefficients. 2) Instead of assuming continuous spatial heterogeneity, we might assume that there exist groups of regions that share the same coefficient. Here, spatial heterogeneity is accommodated by making use of a discrete number of separate classes of regions, Q, where each region has an unknown probability of belonging to some class, and each class has a different coefficient. Thus the groups are homogeneous, with common parameters $\boldsymbol{\beta}_q, q = 1, ..., Q$, for the members of the group, but the groups themselves are different from one another. In this case, we say that the parameters vary across space following a discrete distribution.[7] We extent these modeling strategies by showing how regional scientists can implement them to incorporate spatial heterogeneity in models using variables at both individual and locational level. We also discuss the similarities between these modeling strategies with models discussed above, as well as issues that arise in each of them.

We also perform several simulations studies in order to assess the ability of those models to identify and/or retrieve the true representation of the spatially varying process under a small sample size context, as well as to illustrate how those modeling strategies can be used. To do so, we estimate the models using maximum likelihood and simulated maximum likelihood, depending on nature of $g()$. A binary probit model is used as the true model governing the true data generating process.

The remainder of this paper is organized as follows. Section 2 discusses the continuous and discrete strategies for modeling spatial heterogeneity. The maximum likelihood estimation procedures are presented in Section 3. In Section 4 we show how regional-specific estimates can be obtained. Our simulation experiments are presented in Section 5, and the results analyzed in Section 6. Finally, Section 7 concludes.

2 Modeling Strategies for Spatial Heterogeneity

In this section, we introduce some important notation to develop a general topology of models with spatial heterogeneity. Consider the following structural model

$$\begin{aligned} y^*_{ci} &= \mathbf{x}'_{ci}\boldsymbol{\beta}_c + \epsilon_{ci}, \quad c = 1, ..., C, i = 1, ..., n_c \\ \boldsymbol{\beta}_c &\sim g(\boldsymbol{\beta}_c), \end{aligned} \tag{1}$$

[7] This modeling approach is also known as the latent class (LC) model. See for example Boxall and Adamowicz (2002); Greene and Hensher (2003); Scarpa and Thiene (2005).

where y^*_{ci} is a latent (unobserved) process for individual i in geographical area c (e.g, region, city, country, census track) that we are trying to explain, and ϵ_{ci} is the error term.[8] It is assumed that the vector $(y_{ci}, \mathbf{x}'_{ci}, \boldsymbol{\beta}'_c)'$ is independently and identically distributed. The conditional probability density function of the latent process, $f^*(y_{ic}|\mathbf{x}_{ci}, \boldsymbol{\beta}_c)$, is determined once the nature of the observed y_{ci} and the population pdf of ϵ_{ci} is known. For example, if the observed y_{ci} is binary and ϵ_{ci} is normal distributed, we obtain the traditional probit model. But if ϵ_{ci} is distributed as logistic, then we obtain the binary logit model. For more examples, see Table 1.

The key element in Equation (1) is $\boldsymbol{\beta}_c$. The notation implies that coefficients are associated with region c, representing those region-specific marginal effects on the latent dependent variable. This implies that all individuals located in the same region have the same coefficient, but there exists inter-spatial heterogeneity, i.e., the coefficients vary across regions but not within the region.[9]

However, we do not know how these parameters vary across regions. All we know is that they vary locally with population pdf $g(\boldsymbol{\beta}_c)$. Once $g(\boldsymbol{\beta}_c)$ is specified, we might have a fully parametric or a semi-parametric spatially random parameter model. In the following sections, we discuss the case where $g(\boldsymbol{\beta}_c)$ follows a continuous or discrete distribution, and their respectively implications.

2.1 Continuous Spatial Heterogeneity

Continuous spatial heterogeneity is introduced by assuming that the parameters vary 'randomly' across regions according to some pre-specified 'continuous' distribution. The pdf of the spatially random coefficients in the population is $g(\boldsymbol{\beta}_c|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents, for example, the mean and variance of $\boldsymbol{\beta}_c$. The goal for the researcher is to estimate $\boldsymbol{\theta}$.

2.1.1 Choosing the Distribution of the Spatially Random Coefficient

The distribution of the spatially random parameters can in principle take any shape. The researcher has to choose a priori the distribution according to his beliefs of the domain and boundedness of the coefficients. Therefore, some prior theoretical knowledge of the spatial structure being modeled may lead to a more appropriate choice of the distribution. Below, we will discuss some continuous distributions and their implications.

Normal Distribution: The normal distribution is by far the most widely used distribution for the spatially random parameters. The density of the normal parameter has mean β and standard deviation σ_β, so that $\boldsymbol{\theta} = (\beta, \sigma_\beta)'$. Thus, the coefficient for each region can

[8] Throughout this work we will use location unit, region, or geographical area interchangeably to refer to the subindex c.

[9] We might allow both intra- and inter-spatial heterogeneity by specifying $\boldsymbol{\beta}_{ci} = \boldsymbol{\beta}_c + \boldsymbol{\delta}_{ci}$, where $\boldsymbol{\beta}_c$ is distributed across regions but not over individuals, and $\boldsymbol{\delta}_{ci}$ is distributed over both individuals and regions.

be written as $\beta_c = \beta + \sigma_\beta \eta_c$, where $\eta_c \sim N(0,1)$. An important feature of the normal density is its unboundedness. This implies that every real number has a positive probability of being drawn. Thus, specifying a given coefficient to follow a normal distribution is equivalent to making a priori assumption that there is a proportion of regions with positive coefficient and another proportion with negative ones. As an illustration, Panel A of Figure 1 displays the distribution of the coefficients for all the regions assuming a normal distribution. In this example the population parameters are $\beta = 0.5$ and $\sigma_\beta = 1$. The proportion of regions with positive coefficient is approximately $\Phi(\beta/\sigma_\beta) \cdot 100 \approx 70\%$, which is showed by the gray area. This last fact makes this distribution quite suitable when the researcher assumes that the effect of $x_{k,ci}$ on y^*_{ci} can have both signs depending in the local context of each region. As an example, there exists an extensive literature that uses the city population as a proxy for urbanization economies (see for example Duranton and Puga, 2004). However, in some regions, large population may suggest agglomeration economies, while in other, it may suggest congestion effects (Ali et al., 2007). In other words, β_c for the population density can take positive or negative values across space.

The normal distribution can be also used as an initial exploratory analysis to determine the domain of a coefficient. For example if the estimated parameters are $\widehat{\beta} = 2$ and $\widehat{\sigma}_\beta = 1$, this implies that approximately $\Phi(\widehat{\beta}/\widehat{\sigma}_\beta) \cdot 100 \approx 98\%$ of the regions in the sample have a positive coefficient. Therefore, the researcher may be more inclined to choose a distribution with just positive real domain. One major disadvantage of the normal distribution is that it has infinite tails, which might result in some regions having implausible extreme coefficient values.

Triangular Distribution: This is a continuous probability distribution with probability density function shaped like a triangle (see Panel B of Figure 1). The advantage of this distribution is that it has a definite upper and lower limit, so its tails are shorter than the normal distribution and we avoid extreme coefficients that may result for some regions. The density of a triangular distribution with mean β and spread s_β is zero beyond the range $(\beta - s_\beta, \beta + s_\beta)$, rises linearly from $\beta - s_\beta$ to β, and drops linearly to $\beta + s_\beta$. The parameters $\boldsymbol{\theta} = (\beta, s_\beta)'$ are estimated.

Uniform Distribution: In this case the parameter for each location is equally likely to take on any value in some interval. Suppose that the spread of the uniform distribution is s_β, such that the parameter is uniformly distributed from $\beta - s_\beta$ to $\beta + s_\beta$ as shown in Panel C of Figure 1. Then the parameter can be constructed as $\beta_c = \beta + s_\beta(2u_c - 1)$ where $u_c \sim \text{U}[0,1]$ and the parameters $\boldsymbol{\theta} = (\beta, s_\beta)$ are estimated. The new random draw $(2u_c - 1)$ is distributed as $\text{U}[-1,+1]$, therefore multiplying by s_β gives a uniformly distributed parameter $\pm s_\beta$ (Train, 2009; Hensher and Greene, 2003). The standard deviation of the uniform distribution can be derived from the spread by dividing s_β by $\sqrt{3}$.

Note also that the uniform distribution with a $[0, 1]$ bound is very suitable when there exists spatial heterogeneity in a dummy variable. For this case the restriction is $\beta = s_\beta = 1/2$.

The normal, triangular and uniform distributions permit positive an negative coefficients. However, as we discussed above, the coefficient may present spatial heterogeneity but only in the positive or negative domain. For example, we may be confident that the coefficient for $x_{k,ic}$ is positive for all regions, but still there may exist spatial heterogeneity around the mean. Some widely used distributions with domain in the positive numbers are the log-normal, truncated normal, and Johnson S_b distribution.[10]

Log-normal Distribution: The support of the log-normal distribution is $(0, \infty)$. Formally, the coefficient for each region is specified as $\beta_c = \exp(\beta + \sigma_\beta \eta_c)$ where $\eta_c \sim \mathrm{N}(0, 1)$. The parameters β and σ_β which represent the mean and standard deviation of $\log(\beta_c)$, are estimated. The median, mean, and standard deviation of β_c are $\exp(\beta_c)$, $\exp(\beta_c + \sigma_\beta^2/2)$ and mean $\times \sqrt{\exp(\sigma_\beta^2) - 1}$, respectively (Revelt and Train, 1998; Train, 2009). The main drawback of the log-normal distribution is that it has a very long right-hand tail. This means that we might find regions with unreasonable extreme positive coefficients as shown in Panel E of Figure 1.

Truncated Normal Distribution: The domain of this distribution is $(0, \infty)$ if the normal distribution is truncated below at zero. The parameter for each region is created as $\beta_c = \max(0, \beta + \sigma_{\boldsymbol{\beta}} \eta_c)$ where $\eta_c \sim \mathrm{N}(0, 1)$ with the share below zero massed at zero equal to $\Phi(-\beta/\sigma_\beta)$. Panel D of Figure 1 shows a spatial random parameter distributed as normal with truncation at 0. The distribution was created using the normal distribution of Panel A. Therefore, the share massed at zero is equal to $\Phi(-\beta/\sigma_\beta) \approx 30\%$. A normal distribution truncated at 0 can be useful when the researcher has a priori belief that for some regions the marginal latent effect of the variable is null. The parameters $\boldsymbol{\theta} = (\beta, \sigma_\beta)$ are estimated.

Johnson S_b Distribution: The S_b distribution gives coefficients between 0 and 1, which is also very suitable for dummy variables. The parameter for region c is computed as $\beta_c = \frac{\exp(\beta + \sigma_{\boldsymbol{\beta}} \eta_c)}{1 + \exp(\beta + \sigma_{\boldsymbol{\beta}} \eta_c)}$ where $\eta_c \sim \mathrm{N}(0, 1)$ and the parameters β and σ_β are estimated. The mean, variance and shape are determined by the mean and variance of $\beta + \sigma_{\boldsymbol{\beta}} \eta_c$ which is a normal distributed parameter. If the analyst needs the coefficient to be between 0 and k, then the variable can be multiplied by k. The logic behind this is the following. Since $\beta_c \times x_{ic}$ ranges between $[0, 1]$, then $\beta_c \times k \times x_{ic}$ is the same as $k[0, 1] = [0, k]$. The advantage of the Johnson S_b is that it can be shaped like log-normal distribution, but with thinner tails below the

[10] If some coefficient is expected a priori to be negative for all the regions, one might create the negative of the variable and then include this new variable in the estimation. This 'trick' allows the coefficient to be negative without imposing a sign change in the estimation procedure (Train, 2009).

bound.

For any distribution, all the information about the unobserved spatial heterogeneity is captured by the spread or standard deviation parameter. For example, a significant standard deviation would reveal a spatially non-stationary relationship, and the higher the standard deviation the higher the unobserved spatial heterogeneity in the parameters. Finally, it is worth noting that if only the constant is assumed to be random, then the model is reduced to the random effect model also known as the spatially constant random parameter in the multilevel context (Jones, 1991). If $n_c = 1$ for all C, then the model is reduced to the RCM.

2.1.2 Correlated Spatially Random Parameters and Observed Variations around the Mean

The random parameters can be generalized to include correlation across the parameters. For example, we may be interested in whether regions with greater (lower) β_1 have also greater (lower) values for β_2. If it is true, we would say that both effects are positively correlated across regions. Furthermore, it is likely that the association between y^*_{ci} and x_{ci} is modified by unmeasured regional effects or region-specific unobserved factors. Therefore, by allowing the constant and the slope parameter to be correlated we might be able to identify whether those unobserved factors and the effect of x_{ci} are positively or negatively associated.

As an illustration of the usefulness of the correlated parameters, Wheeler and Tiefelsdorf (2005) raise the awareness of the potential dependencies (correlation) among the local regression coefficients associated with different exogenous variables in the GWR context. They use a GWR approach to explain the white male bladder cancer mortality rates in the 508 States Economic Areas of the United States. Using the population density and smoking as covariates, they find that those regions with high smoking parameter also have a low population density parameter. As they state, the important question is whether this negative correlation is real or an artifact of the statistical method. By allowing the parameters to be explicitly correlated, we are able to test whether the correlation among the parameters is in fact significant.

For simplicity of the notation, consider that the coefficients are distributed across space following a multivariate normal distribution, $\boldsymbol{\beta}_c \sim MVN(\boldsymbol{\beta}, \boldsymbol{\Sigma})$. In this case, the coefficient can be written as:

$$\boldsymbol{\beta}_c = \boldsymbol{\beta} + \mathbf{L}\boldsymbol{\eta}_c,$$

where $\boldsymbol{\eta}_c \sim N(\mathbf{0}, \mathbf{I})$, and $\mathbf{L}$ is the lower-triangular Cholesky factor of $\boldsymbol{\Sigma}$ such that $\mathbf{L}\mathbf{L}' = \text{Var}(\boldsymbol{\beta}_c) = \boldsymbol{\Sigma}$. When the off-diagonal elements of $\mathbf{L}$ are zero the parameters are independently normal distributed. If we assume that the model has only one covariate and the constant,

then the extended form of the spatially random vector is

$$\begin{pmatrix} \alpha_c \\ \beta_c \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \sigma_{\alpha\alpha} & 0 \\ \sigma_{\beta\alpha} & \sigma_{\beta\beta} \end{pmatrix} \begin{pmatrix} \eta_{c\alpha} \\ \eta_{c\beta} \end{pmatrix}$$
$$\boldsymbol{\beta}_c = \boldsymbol{\beta} + \mathbf{L}\boldsymbol{\eta}_c,$$

where:

$$\mathbf{L}\mathbf{L}' = \begin{pmatrix} \sigma_{\alpha\alpha} & 0 \\ \sigma_{\beta\alpha} & \sigma_{\beta\beta} \end{pmatrix} \begin{pmatrix} \sigma_{\alpha\alpha} & \sigma_{\beta\alpha} \\ 0 & \sigma_{\beta\beta} \end{pmatrix} = \begin{pmatrix} \sigma_{\alpha\alpha}^2 & \sigma_{\alpha\alpha}\sigma_{\beta\alpha} \\ \sigma_{\beta\alpha}\sigma_{\alpha\alpha} & \sigma_{\beta\alpha}^2 + \sigma_{\beta\beta}^2 \end{pmatrix} = \boldsymbol{\Sigma}$$

If we need correlated parameters with positive domain, we might create a log-normal distributed parameter. For instance, if we need β_c to be log-normal distributed, then we can transform it in the following way:

$$\beta_c = \exp(\beta + \sigma_{\beta\alpha}\eta_{c\alpha} + \sigma_{\beta\beta}\eta_{c\beta})$$

Observed spatial heterogeneity—or deterministic spatial heterogeneity—can be also accommodated in the random parameters by including region-specific covariates. Specifically, the vector of random coefficient is:

$$\boldsymbol{\beta}_c = \boldsymbol{\beta} + \boldsymbol{\Pi}\mathbf{z}_c + \mathbf{L}\boldsymbol{\eta}_c \tag{2}$$

where $\mathbf{z}_c$ is a set of M characteristics of region c that influences the mean of the spatial random coefficients, and $\boldsymbol{\Pi}$ is a $K \times M$ is a matrix of additional parameters. The conditional mean becomes $\mathbb{E}(\boldsymbol{\beta}_c|\mathbf{z}_c) = \boldsymbol{\beta} + \boldsymbol{\Pi}\mathbf{z}_c$.

The main drawback of this modeling strategy—and any type of spatial heterogeneity in the form of unobserved spatial heterogeneity—is that it assumes that the coefficients are drawn from some univariate or multivariate distribution and no attention is paid to the location of the regions (Fotheringham and Brunsdon, 1999). However, the model in Equation (2) can be very useful to consider regions' location explicitly in the random parameters if $\mathbf{z}_c$ includes any function of the geographical coordinates (u_c, v_c). Thus, if $\mathbf{z}_c = h(\mathbf{u}_c, \mathbf{v}_c)$, where $h()$ is any function, and $\boldsymbol{\eta}_c = 0$, then the model collapses into the Casetti's spatial expansion method. This approach is used in the simulation exercise.

2.2 Discrete Spatial Heterogeneity

Instead of assuming a continuous distribution for the spatially random coefficients, we can assume that they are distributed across space following a discrete distribution. In this case, spatial heterogeneity is accommodated by making use of a discrete number (say Q) of separate classes of regions with different values for the coefficients in each class. The classes can be though as a classification or segmentation of regions, which are homogeneous in terms of the

marginal effect of the variables on the latent process.[11] Formally, the distribution of spatially random parameter is

$$g(\boldsymbol{\beta}_c|\boldsymbol{\theta}_q) = \begin{cases} \boldsymbol{\beta}_1 & \text{with probability } w_{c1} \\ \boldsymbol{\beta}_2 & \text{with probability } w_{c2} \\ \vdots & \vdots \\ \boldsymbol{\beta}_Q & \text{with probability } w_{cQ} \end{cases}, \tag{3}$$

where region c belongs to class q with probability w_{cq}, such that $\sum_q w_{cq} = 1$ and $w_{cq} > 0$. The class assignment probability in Equation (3) is unknown to the analyst. Therefore, the number of classes Q—which is equal to the number of support points—must be chosen a priori by the researcher. The most widely used formulation for w_{cq} is the semi-parametric multinomial logit format[12]

$$w_{cq} = \frac{\exp(\mathbf{h}_c'\boldsymbol{\gamma}_q)}{\sum_{q=1}^{Q}\exp(\mathbf{h}_c'\boldsymbol{\gamma}_q)}; \quad q = 1, ..., Q, \boldsymbol{\gamma}_1 = \mathbf{0},$$

where $\mathbf{h}_c$ represents a vector of regional characteristics that determine the assignment to classes, and $\boldsymbol{\theta}_q = \boldsymbol{\gamma}_q$. The coefficient vector of the first class, $\boldsymbol{\gamma}_1$, is normalized to zero for identification of the model. In the simulation experiments, we let $\mathbf{h}_c$ be a linear function of (u_c, v_c), so that the discrete segmentation of the regions is based on their geographical location. As we will see, this formulation is very useful to detect clusters of regions, where the clusters are in terms similar 'sensitivities'.

Note that one could omit $\mathbf{h}_c$ as determinant of the class assignment probability. The probabilities become:

$$w_q = \frac{\exp(\gamma_q)}{\sum_{q=1}^{Q}\exp(\gamma_q)}; \quad q = 1, ..., Q, \gamma_1 = 0,$$

where γ_q is a constant for class q.

Compared with the continuous spatial heterogeneity model, a discrete spatial heterogeneity approach has the advantage of being a semi-parametric specification, which frees the researcher from potential problems of misspecification in the distribution. In fact, the main disadvantage of the continuous heterogeneity is that the analyst has to choose the distribution of the spatial random parameters a priori, whereas in the discrete heterogeneity case no assumption are made about the shape of the spatial heterogeneity other than the number of classes. As we will discuss in Section 3, another main advantage of this model is

[11] Conceptually, this approach is the same as dividing the regions into Q classes and then conducts separate regression for all the individuals in each class, however it is less efficient.

[12] See for example Boxall and Adamowicz (2002); Greene and Hensher (2003); Scarpa and Thiene (2005).

that it does not require any simulation-based method to estimate the parameters.

3 Likelihood-Based Estimation

Let $\mathbf{y}_c = \{y_{i1}, y_{i2}, ..., y_{in_c}\}$ be the sequence of choices for all individuals in region c, where n_c is the total number of individuals in that region. Assuming that individuals are independent across regions, the joint probability density function, given $\boldsymbol{\beta}_c$, can be written as

$$\Pr(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\beta}_c) = \prod_{i=1}^{n_c} f^*(y_{ic}|\mathbf{x}_{ic}, \boldsymbol{\beta}_c), \tag{4}$$

because, conditional on $\boldsymbol{\beta}_c$, the observations are independent. Since $\boldsymbol{\beta}_c$ is common for individuals living in the region c, within each region individuals are not independent. Thus, the unconditional pdf of $\mathbf{y}_c$ given $\mathbf{X}_c$ will be the weighted average of the conditional probability (4) evaluated over all possible values of $\boldsymbol{\beta}$, which depends on the parameters of the distribution of $\boldsymbol{\beta}_c$. For the discrete and continuous spatial heterogeneity, the unconditional pdf's are respectively

$$P_c(\boldsymbol{\theta}_q) = f(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\theta}_q) = \sum_{q=1}^{Q} w_{iq} \left[\prod_{i=1}^{N_c} f^*(y_{ic}|\mathbf{x}_{ic}, \boldsymbol{\beta}_c, \boldsymbol{\theta}_q) \right], \tag{5}$$

$$P_c(\boldsymbol{\theta}) = f(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\theta}) = \int_{\boldsymbol{\beta}_c} \left[\prod_{i=1}^{N_c} f^*(y_{ic}|\mathbf{x}_{ic}, \boldsymbol{\beta}_c, \boldsymbol{\theta}) \right] g(\boldsymbol{\beta}_c) d\boldsymbol{\beta}_c, \tag{6}$$

In general, the model with discrete spatial heterogeneity and unconditional pdf given in (5) can be easily estimated using maximum likelihood or Expectation-Maximization (EM) algorithm (see for example Ruud, 1991). In the simulation experiment, we estimate the discrete spatial heterogeneity model by maximum likelihood using the Broyden-Frecher-Goldfarb-Shanno (BFGS) maximization algorithm.

Unlike the discrete case, the unconditional probability in expression (6) has no closed form solution, therefore the log-likelihood function is difficult to compute. However, we can simulate this probability and use the simulated maximum likelihood in order to estimate $\boldsymbol{\theta}$ (Gourieroux and Monfort, 1997; Hajivassiliou and Ruud, 1986; Stern, 1997; Train, 2009).[13] In particular, $P_c(\boldsymbol{\theta})$ is approximated by a summation over randomly chosen values of $\boldsymbol{\beta}_c$. For a given value of the parameters $\boldsymbol{\theta}$, a value of $\boldsymbol{\beta}_c$ is drawn from its distribution. Using this draw of $\boldsymbol{\beta}_c$, $P_c(\boldsymbol{\theta})$ from Equation (6) is calculated. This process is repeated for many draws, and

[13] Other methods can be used in order to approximate the integrals. For example, Gauss-Hermite quadrature procedure is another numerical method widely used. However, it has been documented that for models with more than 3 random parameters SML performs better. Bayesian estimation is also suitable for continuous spatial heterogeneity. See for example Hashiguchi and Tanaka (2014).

the average over the draws is the simulated probability. Formally, the simulated probability for region c is

$$\widetilde{P}_c(\boldsymbol{\theta}) = \frac{1}{R}\sum_{r=1}^{R}\prod_{i=1}^{N_c}\widetilde{P}_{icr}(\boldsymbol{\theta}) \tag{7}$$

where $\widetilde{P}_{icr}$ is the probability for individual i in region c evaluated at the rth draw of $\boldsymbol{\beta}_c$, and R is the total number of draws. Then, the simulated log-likelihood function is:

$$\log L_s = \sum_{c=1}^{C}\log\left[\frac{1}{R}\sum_{r=1}^{R}\prod_{i=1}^{N_c}\widetilde{P}_{icr}(\boldsymbol{\theta})\right] \tag{8}$$

Lee (1992), Gourieroux and Monfort (1991) and Hajivassiliou and Ruud (1986) derive the asymptotic distribution of the simulated maximum likelihood (SML) estimator based on smooth probability simulators with the number of draws increasing with sample size. Under regularity conditions, the estimator is consistent and asymptotically normal. When the number of draws, R, rises faster than the square root of the number of observations, the estimator is asymptotically equivalent to the maximum likelihood estimator. It is worth noting that, even though the simulated probability in (7) an unbiased estimate of the true probability, the log of the simulated probability with fixed number of repetitions is not an unbiased estimate of the log of the true probability. This bias in the SML decreases as the number of draws increases (see for example Gourieroux and Monfort, 1997; Revelt and Train, 1998). Like the discrete case, the SML is estimated using the BFGS algorithm in the simulation experiments.

One main limitation of these modeling strategies is that the performance of the maximum likelihood estimators may not be accurate or satisfactory when the number of individual per region is large. The problem is that the log-likelihood function involves the integration or summation over a term involving the product of the probabilities for all the individuals in each location c. Borjas and Sueyoshi (1994) were the first in noticing this problem in the context of the probit model with random effects and using Gauss quadrature. Lee (2000) also gives more insights about this problem. For example, assuming a sample 500 individuals per group—or regions in our case— with a likelihood contribution of 0.5 per observation, Borjas and Sueyoshi (1994) show that the value of the integrand can be as small as $e^{500\times\ln(0.5)} \approx e^{-346.6}$, which is below the existing absolute value for a computer. A consequence of this might be larger standard errors, explosive estimates and/or singular Hessian. In the worst scenario, the computation will 'overflow', that is, it will exceed the computer's capacity to compute the value and the maximization procedure will stop. This issue should be bear in mind when applying these methods.

4 Region-Specific Estimates

In applied literature is very common mapping the region-specific estimates to display the spatial heterogeneity for certain coefficients. This cannot be done using just the distribution of the parameters across regions, $g(\boldsymbol{\beta}_c|\boldsymbol{\theta})$. The population distributions give us just the average affect, β, and the spatial variation around this mean , σ_β, when in fact we would like to know where each region's $\boldsymbol{\beta}_c$ lies in $g(\boldsymbol{\beta}_c|\boldsymbol{\theta})$. We might be able to find the likely location of a given region on the heterogeneity distribution by moving from the conditional to the unconditional distribution (Revelt and Train, 2000; Brunsdon et al., 1999). Using Bayes' theorem we obtain:

$$f(\boldsymbol{\beta}_c|\mathbf{y}_c, \mathbf{X}_c, \boldsymbol{\theta}) = \frac{f(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\beta}_c)g(\boldsymbol{\beta}_c|\boldsymbol{\theta})}{f(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\theta})} = \frac{f(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\beta}_c)g(\boldsymbol{\beta}_c|\boldsymbol{\theta})}{\int_{\boldsymbol{\beta}_c} f(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\beta}_c)g(\boldsymbol{\beta}_c|\boldsymbol{\theta})d\boldsymbol{\beta}_c}, \tag{9}$$

where $f(\boldsymbol{\beta}_c|\mathbf{y}_c, \mathbf{X}_c, \boldsymbol{\theta})$ is the distribution of the regional parameters $\boldsymbol{\beta}_c$ conditional on the sequence of choices of all the individuals in region c, whereas $g(\boldsymbol{\beta}_c|\boldsymbol{\theta})$ is the unconditional distribution. The conditional expectation of $\boldsymbol{\beta}_c$ is given by

$$\bar{\boldsymbol{\beta}}_c = \mathbb{E}\left[\boldsymbol{\beta}_c|\mathbf{y}_c, \mathbf{X}_c, \boldsymbol{\theta}\right] = \frac{\int_{\boldsymbol{\beta}_c} \boldsymbol{\beta}_c f(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\beta}_c)g(\boldsymbol{\beta}_c|\boldsymbol{\theta})d\boldsymbol{\beta}_c}{\int_{\boldsymbol{\beta}_c} f(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\beta}_c)g(\boldsymbol{\beta}_c|\boldsymbol{\theta})d\boldsymbol{\beta}_c} \tag{10}$$

This expectation gives us the conditional mean of the distribution of the spatially random parameter. The simulator of (10) for the continuous and discrete case are, respectively:

$$\widehat{\bar{\boldsymbol{\beta}}_c} = \widehat{\mathbb{E}}\left[\boldsymbol{\beta}_c|\mathbf{y}_c, \mathbf{X}_c, \widehat{\boldsymbol{\theta}}\right] = \frac{\frac{1}{R}\sum_{r=1}^{R} \widehat{\boldsymbol{\beta}}_{cr} \prod_{i=1}^{n_c} f^*(y_{ci}|\mathbf{x}_{ci}, \widehat{\boldsymbol{\beta}}_{cr})}{\frac{1}{R}\sum_{r=1}^{R} \prod_{i=1}^{n_c} f^*(y_{ci}|\mathbf{x}_{ci}, \widehat{\boldsymbol{\beta}}_{cr})} \tag{11}$$

$$\widehat{\bar{\boldsymbol{\beta}}_c} = \widehat{\mathbb{E}}\left[\boldsymbol{\beta}_c|\mathbf{y}_c, \mathbf{X}_c, \widehat{\boldsymbol{\theta}}_q\right] = \frac{\sum_{q=1}^{Q} \widehat{\boldsymbol{\beta}}_q \widehat{w}_{cq} \prod_{i=1}^{n_c} f^*(y_{ci}|\mathbf{x}_{ci}, \widehat{\boldsymbol{\beta}}_q)}{\sum_{q=1}^{Q} \prod_{i=1}^{n_c} f^*(y_{ci}|\mathbf{x}_{ci}, \widehat{\boldsymbol{\beta}}_q)} \tag{12}$$

In addition to estimating the region-specific estimates, we might also like to know whether the parameter for a given region is positive, negative or zero by constructing confidence intervals. In order to construct confidence interval for $\widehat{\bar{\boldsymbol{\beta}}_c}$, we can get an estimator of the conditional variance of $\boldsymbol{\beta}_c$ using the point estimates as follows (Greene, 2012, chapter 15):

$$\widehat{V}_c = \widehat{\mathbb{E}}\left[\boldsymbol{\beta}_c^2|\mathbf{y}_c, \mathbf{X}_c, \widehat{\boldsymbol{\theta}}\right] - \left(\widehat{\mathbb{E}}\left[\boldsymbol{\beta}_c|\mathbf{y}_c, \mathbf{X}_c, \widehat{\boldsymbol{\theta}}\right]\right)^2 \tag{13}$$

An approximate normal-based 95% confidence interval can be then constructed as $\widehat{\bar{\boldsymbol{\beta}}_c} \pm 1.96 \times \widehat{V}_c^{1/2}$. If n_c increases without bound, the estimated conditional variance will approach the estimated variance in the population. It is also expected that $\widehat{\bar{\boldsymbol{\beta}}}_c \to \boldsymbol{\beta}_c$ as $n_c \to \infty$. That is, if we have more information about the choices made by the individuals in each region, then we are in better position to identify where each region coefficient lies on $g(\boldsymbol{\beta}_c)$ (Train, 2009; Revelt and Train, 2000).

5 Simulation Experiments

We use two simulation studies to assess the accuracy of the regression coefficients from models with continuous and discrete spatial heterogeneity, respectively. To do so, we use the binary probit model as the model governing the true data generating process.

Since it is usual to find empirical work using between 50 and 200 spatial units, we mainly focus on the ability of both models to retrieve the true representation of the spatial heterogeneity when the number of regions in the sample is small. Specifically, we address the following issues:

- How do the ML and SML estimates behave when the number of regions is small? In order to give some insights about this, in each simulation experiment we create databases with $C = \{49, 100, 196\}$ regions. Those numbers are chosen such that regions are equally spaced on a square grid of $\sqrt{C} \times \sqrt{C}$. Similar approach is used for example by Wheeler and Calder (2007) and Páez (2005).

- As we discussed in Section 3, it is expected that, if n_c rises without bound, then the conditional means of the parameters, $\widehat{\bar{\boldsymbol{\beta}}}_c$, converges to the true $\boldsymbol{\beta}_c$. Then the question becomes: Given a number of regions, approximately how many individuals per regions do we need in order to get conditional means closer to the true $\boldsymbol{\beta}_c$? In order to address this question, we create databases with $n = \{10, 25, 50, 80, 100\}$ individuals in each regions. We limit the experiments to balanced number of individuals per region for simplicity. Revelt and Train (2000) have addressed this question in the context of the multinomial logit model, where the sub-index n_c correspond to the number of choice situation per individual. We expand on their work by using random coefficients where the means vary across regions according to their geographical location.

- As we mentioned in Section 3, the estimation procedure posits some limitations on the number of individuals per regions due to numerical problems. Thus, we also assess how well the estimation procedures work as the number of individual per region increases.

Given this setting, we have $3 \times 5 = 15$ scenarios in each simulation study. For each scenario we generate $S = 100$ artificial databases (trials). Thus, we estimate $100 \times 3 \times 5 = 1500$ models in each simulation study. In each trial, the explanatory variables and the error terms are simulated, while the true spatially random coefficients are hold fixed in each scenario. This will allow us assessing whether the conditional density of the regional-specific parameters converges to the true population distribution as n_c becomes bigger.[14]

[14] All models are estimated in R (Team, 2015). For the SML we used the Rchoice package (Sarrias, 2014). We coded the ML procedure for the discrete case.

5.1 Experiment 1: Continuous Case

The true latent process is given by:

$$
\begin{aligned}
y^*_{ci} &= \alpha + \beta_1 x_{1,ci} + \beta_{2c} x_{2,ci} + \beta_{3c} x_{3,ci} + \epsilon_{ci} \\
\beta_{2c} &= \beta_2 + \pi_{2u} u_c + \pi_{2v} v_c + \sigma_2 \eta_{2c} \\
\beta_{3c} &= \beta_3 + \pi_{3u} u_c + \pi_{3v} v_c + \sigma_3 \eta_{3c}
\end{aligned}
$$

where $x_{1,ci}, x_{2,ci}$ and $x_{3,ci}$ are independently distributed as normal $\mathrm{N}(0,1)$; u_c and v_c represent the normalized longitude (east and west) and latitude (north and south) coordinates, respectively.[15] The error term is distributed as $\mathrm{N}(0,1)$, so that the model is probit; η_{kc} are independent standard normal variables ($k = 2, 3$), therefore the parameters are normally distributed. The true fixed parameters α and β_1 are both set equal to 1. The true mean values for the regions were set at $\beta_2 = 1$ and $\beta_3 = -1$, and the true values for σ_2 and σ_3 were both set to 1. Finally, the true geographical parameters are $\pi_{2u} = 1$, $\pi_{2v} = -1$, $\pi_{3u} = -1$ and $\pi_{3v} = 1$

The latent variable y^*_{ci} is linked to the observed binary variable y_{ci} by the following rule:

$$
y_{ci} = \begin{cases} 1 & \text{if } y^*_{ci} > 0 \\ 0 & \text{if } y^*_{ci} \leq 0 \end{cases}
$$

All models are estimated using simulated maximum likelihood with 100 Halton draws.[16]

5.2 Experiment 2: Discrete Case

The true latent process is given by:

$$
y^*_{ci} = \beta_{1c} x_{1,ci} + \beta_{2c} x_{2,ci} + \epsilon_{ci}
$$

where

$$
\beta_{1c} = \begin{cases} -1 & \text{for class } q = 1 \\ 0 & \text{for class } q = 2 \\ 0.5 & \text{for class } q = 3 \\ 1 & \text{for class } q = 4 \end{cases}, \quad \beta_{2c} = \begin{cases} -1 & \text{for class } q = 1 \\ -0.5 & \text{for class } q = 2 \\ 0 & \text{for class } q = 3 \\ 1 & \text{for class } q = 4 \end{cases}
$$

and the four classes are crated according to the location of regions in the squared grid. Specifically, we divide the square grid in four areas

[15] The normalization transform the original geographical coordinates so that they range between 0 and 1. See for example Páez (2005).

[16] We also estimated the models using 500 draws, but no changes in the results were observed. We choose 100 draws to keep estimation times manageable.

- Region c belongs to class 1, if $u_c < 0.5$ and $v_c < 0.5$.
- Region c belongs to class 2, if $u_c \geq 0.5$ and $v_c \geq 0.5$.
- Region c belongs to class 3, if $u_c < 0.5$ and $v_c \geq 0.5$.
- Region c belongs to class 4, if $u_c \geq 0.5$ and $v_c < 0.5$.

where u_c and v_c are the normalized geographical coordinates. The latent variable is linked to the observed binary variable as in the continuous case.

6 Results

6.1 Experiment 1: Continuous Case

We start analyzing the results for the continuous spatial heterogeneity experiment. Table 2 shows the average CPU time in seconds that took each trial to convergence, along with the number of actual samples that converged in each scenario. First, we note that CPU times increase approximately linear with C and n. For example, given $n_c = 10$, estimation time increases by a time factor of about 3 when C increases. Second, these results also give us a preliminary analysis of the numerical problems encountered using SML when the number of individuals is greater or equal to 80. The SML algorithm is in general stable when n_c is between 10 and 50. But, numerical problems appear when n_c is greater or equal to 80. For example, when $C = 49$ and $n = 80$ we had to discard 64% of the samples because convergence did not occur or because the SML estimates are unexpectedly too high. Those results agree with those reported by Borjas and Sueyoshi (1994), who pointed out that group sizes over 50 may create significant instabilities.[17] Consequently, we only report the results for $n = 10, 25$ and 50.

Table 3 shows the main statistics for the experiment, which correspond to averages over the S trials. We do not report the statistics for the fixed coefficients. Looking at the estimates, we observe that there are biases. The magnitude of the bias varies according to the type of the parameters. For example, the standard deviations present less bias. In the best scenario ($C = 196, n = 50$) the biases for σ_2 and σ_3 are -0.005 and 0.086, respectively. More striking are the results for the means of the spatially random parameters. They show higher downward bias when the number of regions increases. Regarding to the geographical-coordinates parameters, we observe that the bias is higher than for the means and standard deviations, but it decreases as the number of regions increases. As the number of regions and/or the number of individuals increases, the standard errors of the estimates

[17] Note that in Borjas and Sueyoshi (1994)'s experiment, the sub index n corresponds to the individuals, while c corresponds to group membership.

decrease and are closer to the standard deviations, as expected.

The poor performance of the SML estimates may come from two sources: simulation bias and small sample bias. The former, explained in Section 3, is due to the fact that the log of the simulated probability is not an unbiased estimate of the log of the true probability. It is expected that this bias decreases as the number of draw increases. The latter is expected to be found in the context of SML (See for example Lee, 1992; Gourieroux and Monfort, 1997). In order to disentangle the main source of bias, we have conducted the same experiment as in Table 3, but 1) using 500 Halton draws and 2) increasing $C = 1024$ for $n = 10$ and 25. For the first case, the results (not reported here) are similar to those with 100 Halton draws. The results for the second additional case are presented in Table 4. The general pattern that emerges is that the statistics improve compare with those in Table 3. Except for π_{3u}, all the estimates show very low bias and lower variability compare vis-à-vis with those with C = 196. In view of the foregoing, one interesting observation is that small sample bias seems to dominate the simulation bias.

Now, we look into the regional-specific estimates of β_{2c}. Figure 2 displays the density of the true unconditional distribution $g(\beta_c)$ along with the distribution of the regional conditional means for different settings of n. The conditional means were estimated using Equation (11) and averaged over the S constructed data sets. As discussed in Section 4, if n can increase without bound, then both $\widehat{\bar{\beta}}_c$ and $\widehat{\text{Var}(\bar{\beta}_c)}$ are consistent estimates of β_c, and $\text{Var}(\beta_c)$, respectively.[18] By looking at the densities we can notice that, given a certain number of regions C, the density of the conditional means converges to the conditional population as the number of individuals in each region increases. In other words, as we have more information of the choices made for the individuals in each region, we are in better shape to identify each region-specific estimate. Consider the case when $C = 100$. If $n = 10$, then the distribution of the conditional means shows lower variability than the true distribution of the coefficient. But, as the number of individuals in each region increases, then the bias is reduced and the moments of the distributions overlap better. Another important observation is that, even though there are biases in the model parameters, the conditional estimates are fairly accurate when $n = 50$. We note however that there are some difficulties to correctly estimate the region-specific estimate for extreme negative values, as shown in panel B in Figure 2.

We carried out a similar analysis for confidence intervals (CIs) of the regional-specific estimates for β_{2c}. Figure 3 shows the 95% confidence intervals for the region-specific estimates for $n = 10$, 25 and 50 respectively. The first, second and third row present the results for $C =$ 49, 100 and 196 respectively. The standard errors used to construct the CIs were computed

[18]This poses a dilemma: we need more individuals per regions in order to achieve consistency, but as we discussed above as n increases numerical problems in the optimization procedure may appear.

using Equation (13). The dotted-blue line represents the true β_c. Therefore, the difference between this line and the estimated conditional mean (black points) can be interpreted as the bias for each region. As expected, the CIs are thinner, the coverage improves, and bias decreases as the number of individuals per region increases. For instance we can make better inference about the sign of the regional-specific estimates when $n = 50$ as opposed when $n = 10$ for any number of regions. In the latter case, it is hard to say whether some estimates in between are truly positive or negative due to larger CIs, notwithstanding the true parameters are indeed positive or negative. It seems also that what matter the most for better inference is n: if we hold n fixed and look across the graphs for different C the pattern in terms of CIs and bias is almost the same. Finally, one can also observe that when the true regional-estimates have very extreme values are more difficult to estimate and show greater bias.

6.2 Experiment 2: Discrete Case

Table 5 presents the average CPU time and the number of trials that converged for the second experiment. The first thing to notice is that the numerical instabilities in the discrete case are not as severe as in the continuous case. For example, in the worse scenario $(C = 100, n = 80)$ only 34% of the trials had to be discarded. Another important observation is the cost in terms of computation time. Since the maximum likelihood optimization under the discrete case does not require simulation, the computation time is much lower than the continuous case.

Results for the discrete spatial heterogeneity experiment are presented in Table 6 for β_1 and Table 7 for β_2. Given the number of estimation and parameters, it is almost impossible to comment in detail each coefficient for each class. Given this restriction, we discuss the results for β_1, and the estimates for the geographical coordinates are not reported. The results reveal that the estimates of the parameters are fairly accurate, notwithstanding a relative small size of sample. In general, the bias is reduced as the number of region and/or individual increases. For example, $\beta_{1,q=1}$ is off by 3% when $C = 49$, but this bias is about 0.9% when C increases to 149, holding fixed the number of individuals at 50. Nevertheless there are some exceptions to the reduction of bias, especially when C = 49. For instance, the parameters' bias for classes 2, 3 and 4 increases when the number of individuals increases from 80 to 100. This pattern is not longer observed when the number of regions increases. Greater efficiency is evidenced in the RMSE results, which shows much lower variation across the runs when the number of individuals and/or regions increase.

Figure 4 display the kernel estimates for the conditional means of β_1. As expected, the precision of the conditional means falls when there are fewer individuals per regions. On the other hand, with 100 individuals per region the distribution of the conditional means overlaps almost perfectly the true discrete distribution of the parameters, especially when the number

of regions in the sample increases. Unlike the continuous case, the bias is almost nonnegligible.

Assuming a discrete distribution for representing the spatial non-stationarity, and allowing the membership probability to depend on the geographical coordinates is also very convenient to detect cluster of regions, where the clusters are in terms of latent marginal effects. To visualize this, Figure 5 shows again the estimates of the conditional means but plotted on the spatial grid for C = 196. Panel A shows the true spatial non-stationary pattern created by the class assignment presented in Section 5.2. For instance, the southwest cluster (or class 1) is characterized by regions with a coefficient for $\beta_1 = -1$; the northwest cluster is characterized by regions where the variable has no effect on the dependent variable; and so on. We can observe that the spatial pattern of the parameter increasingly resembles the true map surface as the number of individuals per region increases.

As in the continuous case, we have also plotted the CIs for β_1 under the discrete case. The results are presented in Figure 6. Again the first, second and third row present the results for $C = 49$, 100 and 196 respectively; and the dotted-blue line represents the true discrete β_1 for the classes. The CIs are larger when $n = 10$. For the three different values of C, we can observe that the intervals for some regions contain zero and positive values when the true parameter is for example -1. In the same vein, some regions have confidence intervals that contain negative values when the true parameter is indeed 0.5. As n increases the CIs become rapidly thinner and their precision improves.

7 Discussion and Conclusion

This paper contributes to the literature of spatial econometric models that deal with spatially non-stationary process by examining continuous and discrete unobserved spatial heterogeneity. These two models have been widely used in discrete choice modeling, however we show how these models can be implemented in order to capture and model spatial heterogeneity. One of the main advantage of the two models is that allow to the analyst to include variables at the individual level, which mitigate the ecological fallacy problem.

In both models, spatial heterogeneity is represented by some distribution $g(\boldsymbol{\beta}_c)$. In the model with continuous spatial heterogeneity, $g(\boldsymbol{\beta}_c)$ can take any continuous shape, and the analyst must choose the distribution a priori. The choice of the distribution may be guided by theoretical reasons regarding to the domain and bound of the coefficients. We discussed also some extensions that can be useful to take into consideration the geographical location of the regions, as well as the spatial correlation of the parameters. In the discrete case, spatial heterogeneity is accommodated by making use of a discrete number of separate classes of regions, thus $g(\boldsymbol{\beta}_c)$ is discrete and modeled in a semi-parametric way. We show how the discrete distribution can be useful to detect cluster of regions in terms of 'sensitivities when

the probability of the class assignment includes the geographical coordinates.

Although both models have very appealing features, there are some differences between them. In terms of estimation, the probability for each region has not closed form solution when $g(\boldsymbol{\beta}_c)$ is continuous. Therefore, we need to simulate this probability and estimate the parameters using SML or a Bayesian approach, which can be very costly in terms of computational time. When $g(\boldsymbol{\beta}_c)$ is discrete, the probability does have a closed form and no simulation is required. Another difference is that the discrete case has the advantage of being a semiparametric specification, which frees the analyst from potential problems of misspecification in the distribution of spatial heterogeneity. In fact, the only sensitive choice in the discrete case is the number of support points that is equal to the number of classes. The main disadvantage of the discrete case is the proliferation of parameters, which increases linearly with the number of classes.

We also conducted simulation experiments to analyze the ability of both approaches to retrieve the true representation of the spatially varying process using small sample sizes. The main finding is that the data requirement of the continuous case is substantial: our results show that models with continuous spatial heterogeneity show greater bias than the discrete case in small samples. The bias of the SML method achieves 'acceptable levels' when the number of regions is around 1000. This result recommends caution, especially if policy implications are based on the result of continuous spatial heterogeneity with small sample.

Regarding to the regional-specific estimates, we found that in both cases the precision to identify each regional parameter improves as the number of individuals per region increases. Nevertheless, the discrete case is able to retrieve the true spatial heterogeneity surface with lower bias and better coverage when compared with the continuous spatial heterogeneity. All these findings tend to favor the discrete case, at least for small samples.

This work can be extended in different ways. First, one of the main concern and limitation of both models is that the estimation requires computing the product of the probabilities for all individuals in a given region. Thus, if the number of individuals is too high, the estimation method may run into numerical difficulties. To overcome this problem some of the two methods proposed by Lee (2000) can be studied under the spatial context. These methods alleviate the numerical problems by interchanging the inner product with the outer summation. Another possible extension is to study both models with small and large samples using Bayesian and EM algorithms. Finally, empirical applications and for both models are needed in order to understand their strengths and weaknesses for estimating models with locally varying coefficients.

References

Aitkin, M. (1996). A General Maximum Likelihood Analysis of Overdispersion in Generalized Linear Models. *Statistics and Computing*, 6(3):251–262.

Ali, K., Partridge, M. D., and Olfert, M. R. (2007). Can Geographically Weighted Regressions Improve Regional Analysis and Policy Making? *International Regional Science Review*, 30(3):300–329.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, volume 4. Springer.

Anselin, L. (2002). Under the Hood Issues in the Specification and Interpretation of Spatial Regression Models. *Agricultural Economics*, 27(3):247–267.

Borjas, G. J. and Sueyoshi, G. T. (1994). A Two-stage Estimator for Probit Models with Structural Group Effects. *Journal of Econometrics*, 64(1):165–182.

Boxall, P. C. and Adamowicz, W. L. (2002). Understanding Heterogeneous Preferences in Random Utility Models: A Latent Class Approach. *Environmental and Resource Economics*, 23(4):421–446.

Brown, L. A. and Jones, J. P. (1985). Spatial Variation in Migration Processes and Development: A Costa Rican Example of Conventional Modeling Augmented by the Expansion Method. *Demography*, 22(3):327–352.

Brunsdon, C., Aitkin, M., Fotheringham, S., and Charlton, M. (1999). A Comparsion of Random Coefficient Modelling and Geographically Weighted Regression for Spatially Non-stationary Regression Problems. *Geographical and Environmental Modelling*, 3:47–62.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. (1998a). Spatial Nonstationarity and Autoregressive Models. *Environment and Planning A*, 30(6):957–973.

Brunsdon, C., Fotheringham, S., and Charlton, M. (1998b). Geographically Weighted Regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(3):431–443.

Casetti, E. (1972). Generating Models by the Expansion Method: Applications to Geographical Research. *Geographical Analysis*, 4(1):81–91.

Casetti, E. (1997). The Expansion Method, Mathematical Modeling, and Spatial Econometrics. *International Regional Science Review*, 20(1-2):9–33.

Casetti, E. and Jones III, J. P. (2003). *Applications of the Expansion Method*. Routledge.

Charlton, M. and Brunsdon, C. (1997). Two Techniques for Exploring Non-stationarity in Geographical Data. *Geographical Systems*, 4:59–82.

Duncan, C. and Jones, K. (2000). Using Multilevel Models to Model Heterogeneity: Potential and Pitfalls. *Geographical Analysis*, 32(4):279–305.

Duranton, G. and Puga, D. (2004). Micro-Foundations of Urban Agglomeration Economies. *Handbook of Regional and Urban Economics*, 4:2063–2117.

Fotheringham, S, A., Charlton, M., and Brunsdon, C. (1996). The Geography of Parameter Space: An Investigation of Spatial Non-stationarity. *International Journal of Geographical Information Systems*, 10(5):605–627.

Fotheringham, A. S. (1997). Trends in Quantitative Methods I: Stressing the Local. *Progress in Human Geography*, 21:88–96.

Fotheringham, A. S. and Brunsdon, C. (1999). Local Forms of Spatial Analysis. *Geographical Analysis*, 31(4):340–358.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2003). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2009). Geographically Weighted Regression. *The Sage handbook of spatial analysis*, pages 243–254.

Goldstein, H. (1987). *Multilevel Models in Education and Social Research*. Oxford University Press.

Gourieroux, C. and Monfort, A. (1991). Simulation Based Inference in Models with Heterogeneity. *Annales d'Economie et de Statistique*, (20-21):69–107.

Gourieroux, C. and Monfort, A. (1997). *Simulation-Based Econometric Methods*. Oxford University Press.

Greene, W. H. (2012). *Econometric Analysis*. Prentice Hall, 7 edition.

Greene, W. H. and Hensher, D. A. (2003). A Latent Class Model for Discrete Choice Analysis: Contrasts With Mixed Logit. *Transportation Research Part B: Methodological*, 37(8):681–698.

Hajivassiliou, V. A. and Ruud, P. A. (1986). Classical Estimation Methods for LDV Models Using Simulation. *Handbook of Econometrics*, 4:2383–2441.

Hashiguchi, Y. and Tanaka, K. (2014). Agglomeration and Firm-Level productivity: A Bayesian Spatial Approach. *Papers in Regional Science*.

Hensher, D. A. and Greene, W. H. (2003). The Mixed Logit Model: The State of Practice. *Transportation*, 30(2):133–176.

Jetz, W., Rahbek, C., and Lichstein, J. W. (2005). Local and Global Approaches to Spatial Data Analysis in Ecology. *Global Ecology and Biogeography*, 14(1):97–98.

Jones, K. (1991). Specifying and Estimating Multi-level Models for Geographical Research. *Transactions of the Institute of British Geographers*, pages 148–159.

Kochanowski, P. (1990). The Expansion Method as a Tool of Regional Analysis. *Regional Science Perspectives*, 20(2):52–66.

Lee, L.-F. (1992). On Efficiency of Methods of Simulated Moments and Maximum Simulated Likelihood Estimation of Discrete Response Models. *Econometric Theory*, 8(04):518–552.

Lee, L.-f. (2000). A Numerically Stable Quadrature Procedure for the One-factor Random-Component Discrete Choice Model. *Journal of Econometrics*, 95(1):117–129.

Lloyd, C. D. (2010). *Local Models for Spatial Analysis*. CRC Press.

McMillen, D. P. (1996). One Hundred Fifty Years of Land Values in Chicago: A Nonparametric Approach. *Journal of Urban Economics*, 40(1):100–124.

Páez, A. (2005). Local Analysis of Spatial Relationships: A Comparison of GWR and the Expansion Method. In *Computational Science and Its Applications–ICCSA 2005*, pages 162–172. Springer.

Páez, A., Farber, S., and Wheeler, D. (2011). A Simulation-based Study of Geographically Weighted Regression as a Method for Investigating Spatially Varying Relationships. *Environment and Planning-Part A*, 43(12):2992.

Peeters, L. and Chasco, C. (2006). Ecological Inference and Spatial Heterogeneity: An Entropy-Based Distributionally Weighted Regression Approach. *Papers in Regional Science*, 85(2):257–276.

Revelt, D. and Train, K. (1998). Mixed Logit With Repeated Choices: Households' Choices of Appliance Efficiency Level. *Review of Economics and Statistics*, 80(4):647–657.

Revelt, D. and Train, K. (2000). Customer-Specific Taste Parameters and Mixed Logit: Households' Choice of Electricity Supplier. Working paper, Department of Economics, UCB.

Robinson, W. (1950). Ecological Correlations and the Behavior of Individuals. *American Sociological Review*, 15(3).

Ruud, P. A. (1991). Extensions of Estimation Methods Using the EM Algorithm. *Journal of Econometrics*, 49(3):305–341.

Sarrias, M. (2014). *Rchoice: Discrete Choice (Binary, Poisson and Ordered) Models with Random Parameters*. R package version 0.2.

Scarpa, R. and Thiene, M. (2005). Destination Choice Models for Rock Climbing in the Northeastern Alps: A Latent-class Approach Based on Intensity of Preferences. *Land Economics*, 81(3):426–444.

Stern, S. (1997). Simulation-Based Estimation. *Journal of Economic Literature*, 35(4):2006–2039.

Swamy, P. (1971). *Statistical Inference in Random Coefficient Regression Models*. Number 55. Springer Berlin.

Team, R. C. (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Train, K. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press.

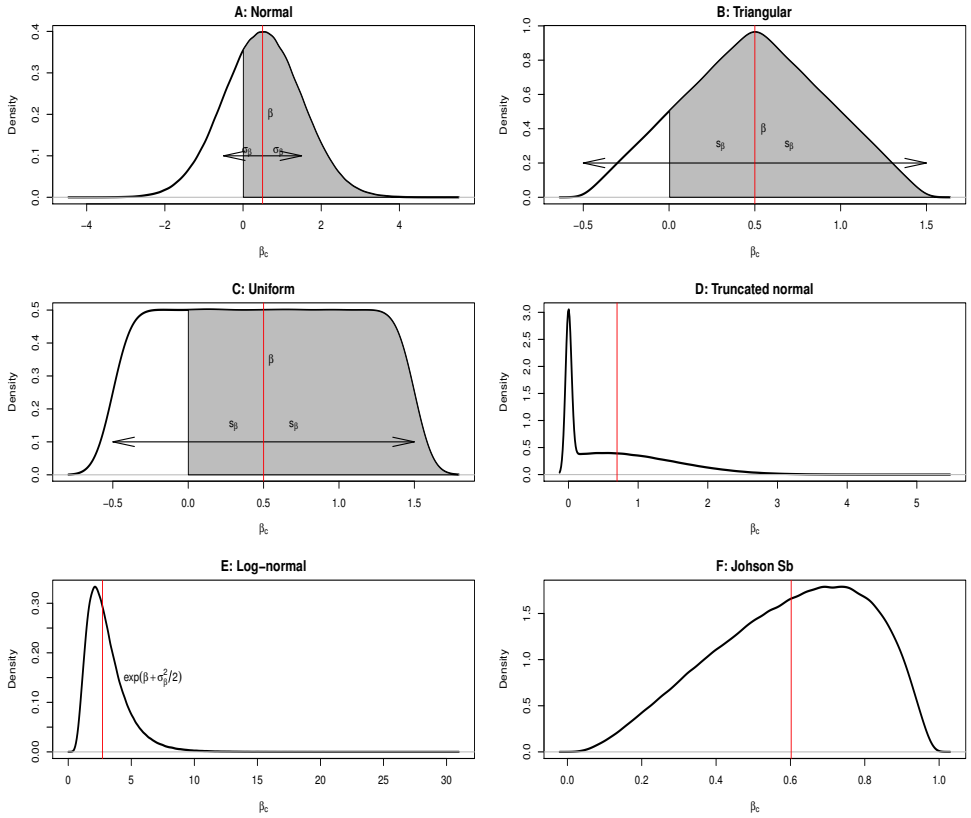
Wheeler, D. and Tiefelsdorf, M. (2005). Multicollinearity and Correlation among Local Regression Coefficients in Geographically Weighted Regression. *Journal of Geographical Systems*, 7(2):161–187.

Wheeler, D. C. and Calder, C. A. (2007). An Assessment of Coefficient Accuracy in Linear Regression Models with Spatially Varying Coefficients. *Journal of Geographical Systems*, 9(2):145–166.

Withers, S. D. (2001). Quantitative Methods: Advancement in Ecological Inference. *Progress in Human Geography*, 25(1):87–96.

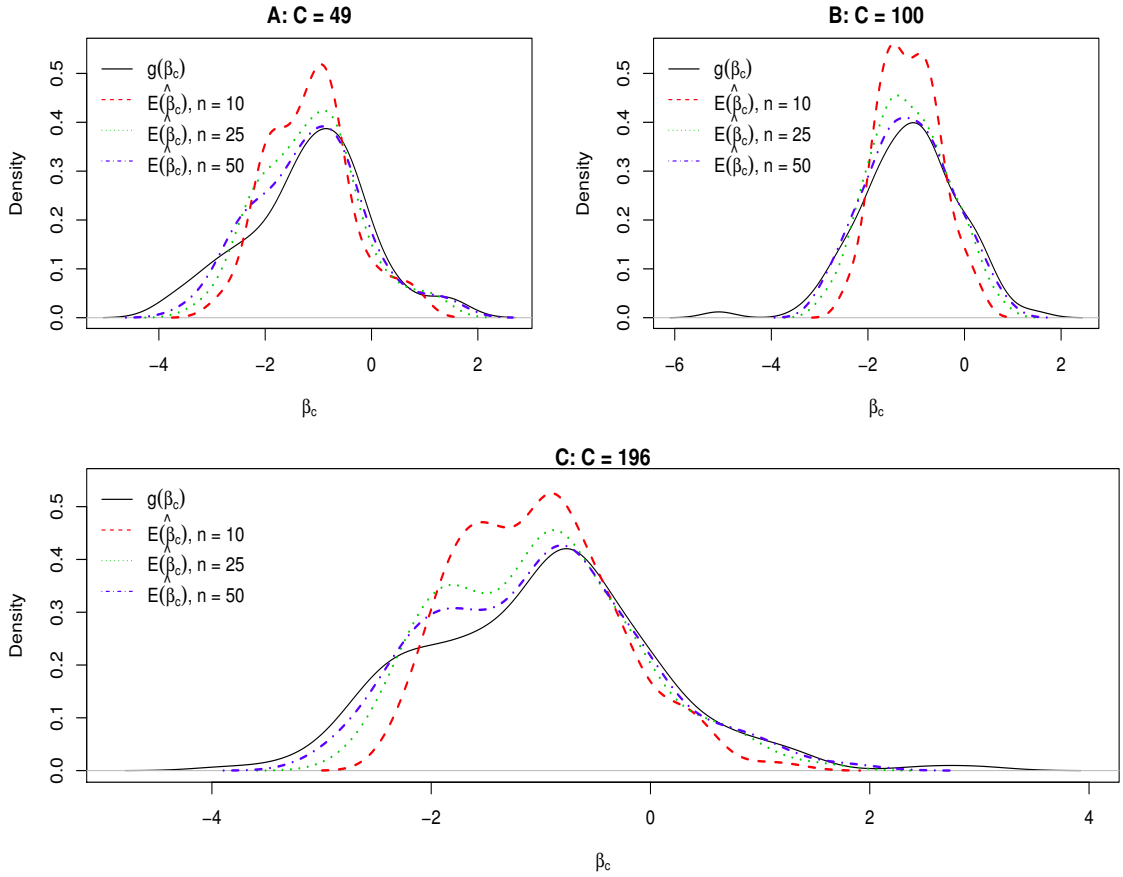
A Figures

Figure 1: Continuous Distributions for the Spatially Random Parameter



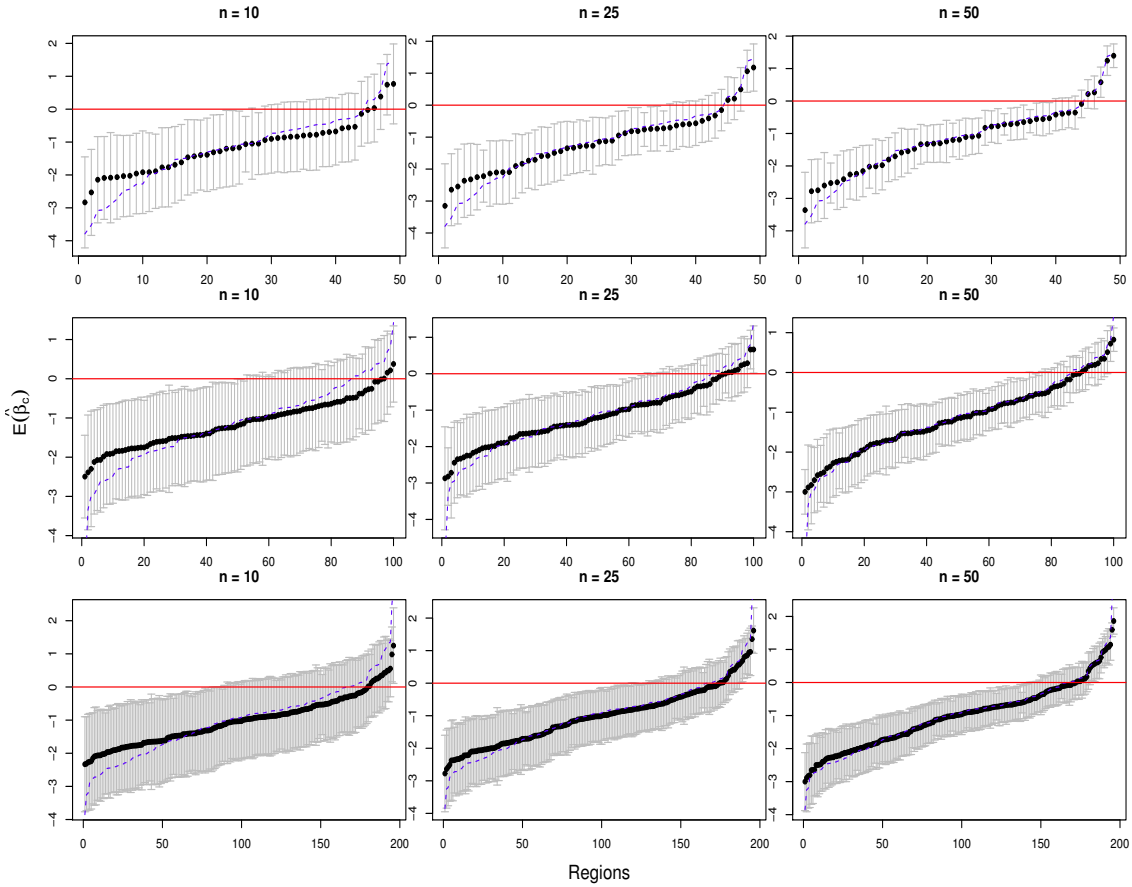
Notes: All the distributions are kernel estimates. The number of regions used for the draws is $C = 1,000,000$.

Figure 2: Conditional Estimates of β_{2c} Under Continuous Spatial Heterogeneity



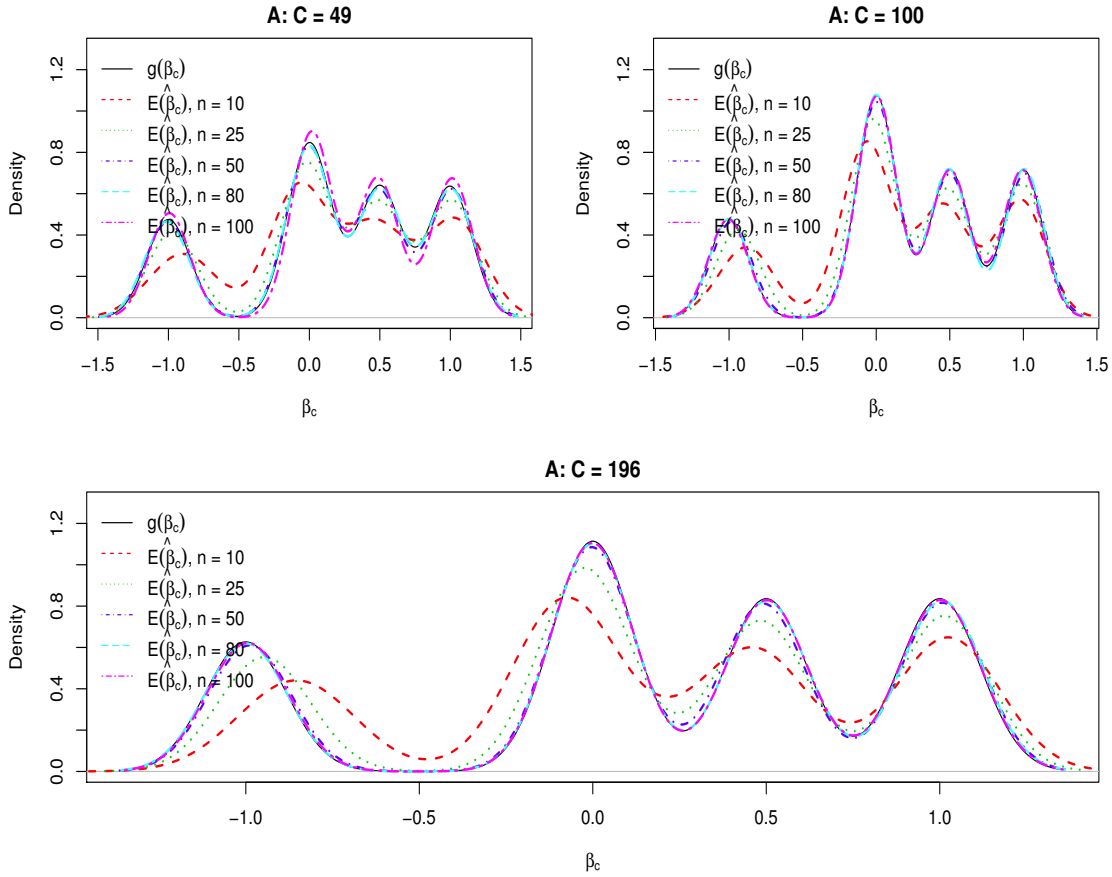
Notes: All the distributions are kernel estimates. The function $g(\boldsymbol{\beta}_c)$ corresponds to the true distribution of the spatially random parameter, which is hold fixed during the simulation experiments. Each point in the distribution of the conditional means corresponds to the average over the S samples for each region in each scenario.

Figure 3: Confident Intervals for Continuous Spatial Heterogeneity: β_{2c}



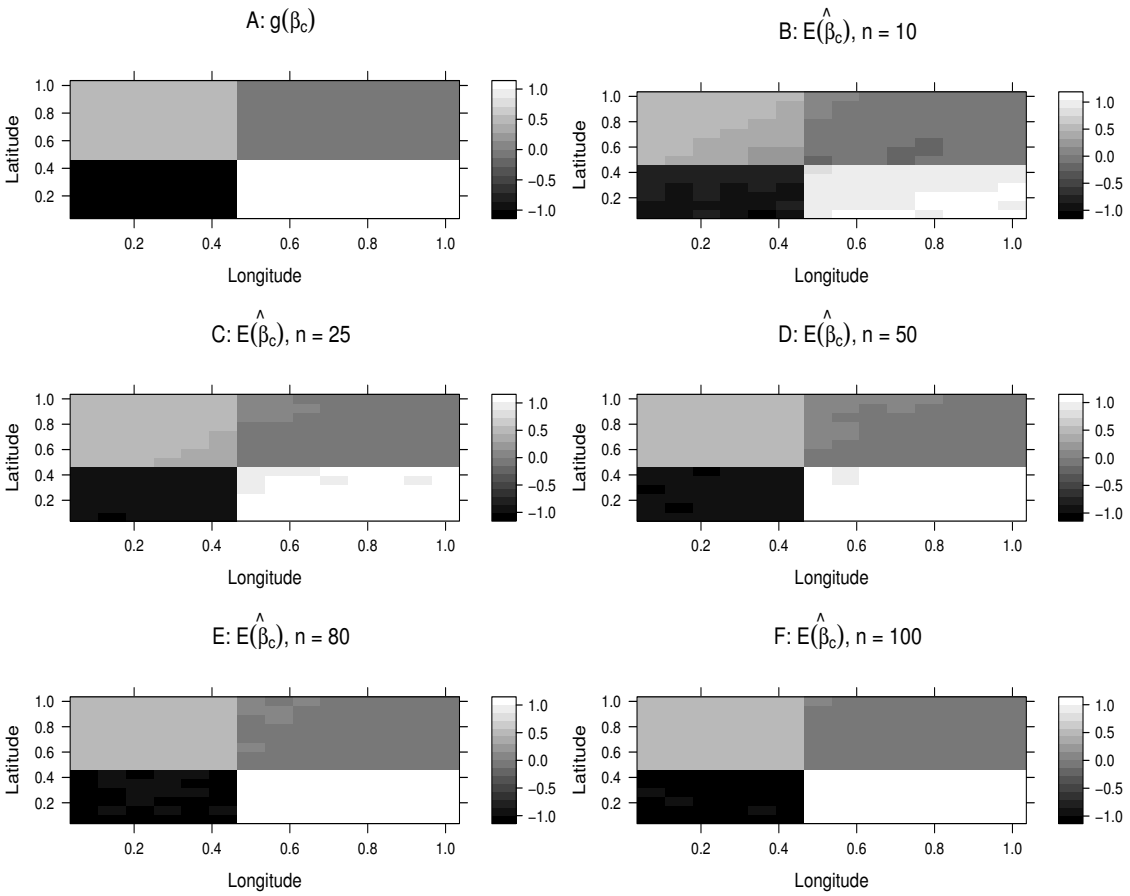
Notes: The blue-dashed line corresponds to the true regional parameter β_{2c}, which follows a normal distribution. The conditional means and standard errors corresponds to the average over the S samples for each region in each scenario.

Figure 4: Conditional Estimates of β_{1c} Under Discrete Spatial Heterogeneity



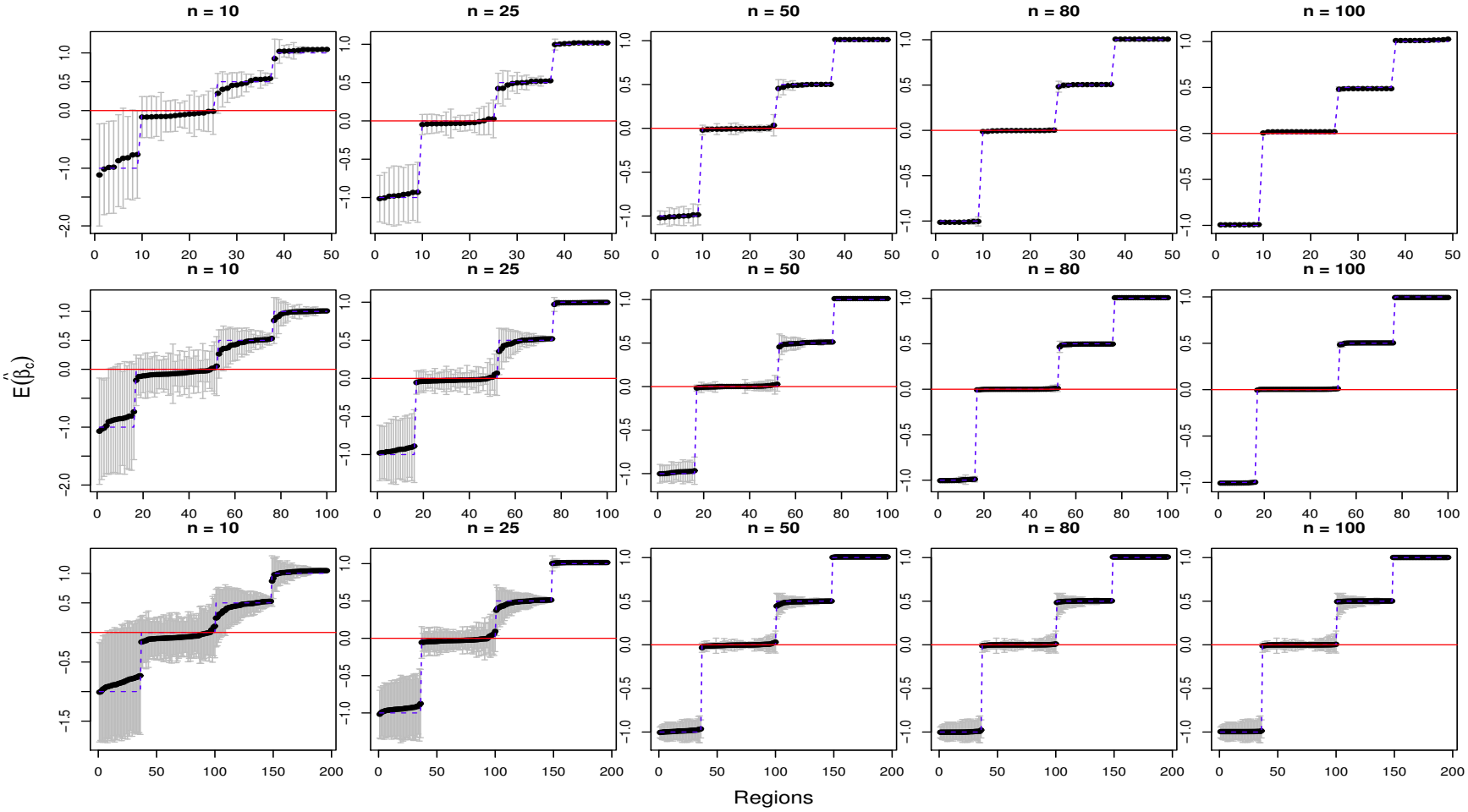
Notes: All the distributions are kernel estimates. The function $g(\boldsymbol{\beta}_c)$ corresponds to the true distribution of the spatially random parameter, which is hold fixed during the simulation experiments. Each point in the distribution of the conditional means corresponds to the average over the S samples for each region in each scenario.

Figure 5: Conditional Estimates Discrete Spatial Heterogeneity: β_{1c} and $C = 196$



Notes: The conditional estimates for each region are plotted in a 14×14 grid. Each square in the grid correspond to a region. The conditional means corresponds to the average over the S samples for each region in each scenario.

Figure 6: Confidence Intervals for Discrete Spatial Heterogeneity: β_{1c}



Notes: The blue-dashed line corresponds to the true regional parameter β_{1c} which follows a discrete distribution. The conditional means and standard errors corresponds to the average over the S samples for each region in each scenario.

B Tables

Table 1: Latent PDFs for Different Models

Model	*Latent PDF*
Linear Model	$f^*(y_{ic}\|\mathbf{x}_{ci},\boldsymbol{\beta}_c) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}\left(y_{ci}-\mathbf{x}'_{ci}\boldsymbol{\beta}_c\right)\right]$
Probit Model	$f^*(y_{ic}\|\mathbf{x}_{ci},\boldsymbol{\beta}_c) = \left[\Phi\left(\mathbf{x}'_{ci}\boldsymbol{\beta}_c\right)\right]^{y_{ci}}\left[1-\Phi\left(\mathbf{x}'_{ci}\boldsymbol{\beta}_c\right)\right]^{1-y_{ci}}$
Multinomial Logit Model	$f^*(y_{ic}\|\mathbf{x}_{cij},\boldsymbol{\beta}_c) = \prod_{j=1}^{J}\left[\frac{\exp\left(\mathbf{x}'_{cij}\boldsymbol{\beta}_c\right)}{\sum_{j=1}^{J}\exp\left(\mathbf{x}'_{cij}\boldsymbol{\beta}_c\right)}\right]^{y_{cij}}$
Ordered Probit Model	$f^*(y_{ic}\|\mathbf{x}_{ci},\boldsymbol{\beta}_c) = \prod_{j=1}^{J}\left[\Phi(k_j-\mathbf{x}'_{ci}\boldsymbol{\beta}_c)-\Phi(k_{j-1}-\mathbf{x}'_{ci}\boldsymbol{\beta}_c)\right]^{y_{cij}}$
Poisson Model	$f^*(y_{ic}\|\mathbf{x}_{ci},\boldsymbol{\beta}_c) = \frac{1}{y_{ci}!}\exp\left[-\exp\left(\mathbf{x}'_{ci}\boldsymbol{\beta}_c\right)\right]\exp\left(\mathbf{x}'_{ci}\boldsymbol{\beta}_c\right)^{y_{ci}}$

Table 2: CPU Time and Trails that Converged for Experiment 1

	Number of Regions					
	C = 49		***C = 100***		***C = 196***	
# Individuals	CPU Time	S	CPU Time	S	CPU Time	S
n=10	11	100	24	100	71	100
n=25	21	100	56	100	161	100
n=50	49	99	138	100	263	100
n=80	108	36	305	67	715	9
n=100	108	34	212	64	404	53

Note: CPU time is the average time (over S) in seconds that took each trial to converge. S corresponds to the total trials that converged. n refers to the total number of individuals per region.

Table 3: Simulation Results for Continuous Spatial Heterogeneity

		C = 49					**C = 100**					**C = 196**				
$\boldsymbol{\theta}_0$	n	**Bias**	**SD**	**RMSE**	**Coverage**	**SE**	**Bias**	**SD**	**RMSE**	**Coverage**	**SE**	**Bias**	**SD**	**RMSE**	**Coverage**	**SE**
	10	-0.110	0.378	0.392	0.990	0.522	-0.144	0.235	0.275	1.000	0.355	-0.230	0.193	0.300	0.950	0.258
$\beta_2 = -1$	25	-0.031	0.254	0.255	1.000	0.439	-0.111	0.166	0.199	1.000	0.292	-0.179	0.149	0.233	0.940	0.216
	50	-0.010	0.224	0.223	0.990	0.374	-0.074	0.196	0.209	0.980	0.234	-0.153	0.172	0.230	0.870	0.189
	10	-0.161	0.199	0.255	0.850	0.215	-0.125	0.160	0.202	0.780	0.149	-0.054	0.089	0.104	0.920	0.104
$\sigma_2 = 1$	25	-0.129	0.110	0.169	0.790	0.138	-0.078	0.087	0.116	0.890	0.098	-0.047	0.046	0.066	0.980	0.068
	50	-0.096	0.104	0.141	0.778	0.123	-0.057	0.076	0.095	0.850	0.076	-0.005	0.056	0.056	0.940	0.059
	10	-0.118	0.355	0.372	1.000	0.543	-0.285	0.232	0.366	0.970	0.397	-0.277	0.153	0.316	0.940	0.263
$\beta_3 = 1$	25	-0.248	0.256	0.355	1.000	0.461	-0.328	0.193	0.380	0.940	0.343	-0.339	0.134	0.364	0.790	0.225
	50	-0.323	0.291	0.433	0.949	0.454	-0.379	0.225	0.440	0.750	0.272	-0.375	0.156	0.406	0.520	0.207
	10	-0.047	0.180	0.185	0.950	0.213	0.053	0.162	0.169	0.940	0.159	0.025	0.096	0.098	0.970	0.107
$\sigma_3 = 1$	25	-0.031	0.121	0.124	0.930	0.141	0.086	0.087	0.122	0.940	0.109	0.038	0.054	0.066	0.980	0.074
	50	0.106	0.104	0.148	0.949	0.123	0.106	0.081	0.133	0.840	0.086	0.072	0.054	0.090	0.820	0.058
	10	0.522	0.480	0.708	0.960	0.622	-0.046	0.277	0.279	1.000	0.426	0.249	0.208	0.324	0.970	0.317
$\pi_{2u} = 1$	25	0.463	0.309	0.556	0.960	0.503	0.028	0.195	0.196	1.000	0.350	0.251	0.182	0.310	0.930	0.262
	50	0.530	0.323	0.620	0.818	0.443	0.094	0.247	0.263	0.970	0.291	0.338	0.209	0.397	0.650	0.228
	10	-0.684	0.421	0.803	0.910	0.649	0.548	0.278	0.613	0.860	0.468	0.368	0.196	0.417	0.930	0.326
$\pi_{3u} = -1$	25	-0.573	0.322	0.657	0.900	0.520	0.619	0.217	0.656	0.790	0.401	0.428	0.155	0.455	0.740	0.275
	50	-0.533	0.340	0.632	0.879	0.454	0.710	0.275	0.761	0.400	0.314	0.453	0.162	0.481	0.530	0.246
	10	-0.602	0.418	0.732	0.950	0.626	0.066	0.317	0.322	1.000	0.429	0.122	0.205	0.238	0.970	0.315
$\pi_{2v} = -1$	25	-0.664	0.316	0.734	0.880	0.511	-0.064	0.217	0.225	1.000	0.363	0.058	0.165	0.174	0.990	0.270
	50	-0.728	0.281	0.780	0.606	0.458	-0.185	0.250	0.310	0.950	0.308	-0.017	0.179	0.179	0.990	0.232
	10	0.413	0.462	0.618	0.990	0.658	0.181	0.300	0.349	0.990	0.485	-0.078	0.218	0.231	0.990	0.329
$\pi_{3v} = 1$	25	0.459	0.328	0.563	0.980	0.551	0.182	0.237	0.298	1.000	0.415	-0.067	0.167	0.179	1.000	0.284
	50	0.419	0.307	0.519	0.970	0.547	0.169	0.324	0.364	0.940	0.338	-0.067	0.218	0.227	0.960	0.245

Note: Bias is $\sum_{s=1}^{S}(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_s)/S$. Standard deviation: SD $= \sqrt{\sum_{s=1}^{S}(\widehat{\boldsymbol{\theta}}_s - \bar{\widehat{\boldsymbol{\theta}}})^2/(S-1)}$. Coverage is the proportion of CIs that contain the true parameter. RMSE $=\sqrt{\sum_{s=1}^{S}(\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_0)^2/S}$. SE correspond to the mean over S of the asymptotic standard error. All models correspond to a binary probit model estimated by SML using 100 Halton Draws. S corresponds to the actual trials that converged (See Table 2).

Table 4: Simulation Results for Continuous Spatial Heterogeneity: $C = 1024$

θ_0	n	**Bias**	**SD**	**RMSE**	**Coverage**	**SE**	**CPU time**	**S**
$\beta_2 = -1$	10	-0.036	0.074	0.082	0.950	0.111	1633	100
	25	-0.019	0.054	0.057	1.000	0.092	3188	100
$\sigma_2 = 1$	10	-0.005	0.044	0.044	0.920	0.047	1633	100
	25	0.000	0.024	0.024	0.990	0.032	3188	100
$\beta_3 = 1$	10	0.098	0.076	0.124	0.940	0.112	1633	100
	25	0.076	0.060	0.096	0.980	0.094	3188	100
$\sigma_3 = 1$	10	0.007	0.040	0.041	0.970	0.047	1633	100
	25	0.025	0.026	0.036	0.920	0.033	3188	100
$\pi_{2u} = 1$	10	0.020	0.089	0.090	0.970	0.141	1633	100
	25	-0.004	0.066	0.065	0.990	0.118	3188	100
$\pi_{3u} = -1$	10	-0.236	0.099	0.256	0.930	0.143	1633	100
	25	-0.184	0.074	0.198	0.800	0.121	3188	100
$\pi_{2v} = -1$	10	0.092	0.092	0.129	0.970	0.141	1633	100
	25	0.102	0.067	0.122	0.980	0.120	3188	100
$\pi_{3v} = 1$	10	-0.050	0.094	0.107	0.990	0.141	1633	100
	25	-0.087	0.086	0.122	0.970	0.120	3188	100

Note: Bias is $\sum_{s=1}^{S}(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_s)/S$. Standard deviation: SD $= \sqrt{\sum_{s=1}^{S}(\widehat{\boldsymbol{\theta}}_s - \bar{\widehat{\boldsymbol{\theta}}})^2/(S-1)}$. Coverage is the proportion of CIs that contain the true parameter. RMSE $=\sqrt{\sum_{s=1}^{S}(\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_0)^2/S}$. SE correspond to the mean over S of the asymptotic standard error. All models correspond to a binary probit model estimated by SML using 100 Halton Draws. S corresponds to the actual trials that converged.

Table 5: CPU Time and Trials that Converged for Experiment 2

	Number of Regions					
	$C =$ **49**		$C =$ **100**		$C =$ **196**	
# Individuals	CPU Time	S	CPU Time	S	CPU Time	S
10	2	88	4	98	7	99
25	4	100	9	100	16	100
50	7	100	15	100	33	100
80	13	100	38	76	42	100
100	22	84	45	100	76	96

Note: CPU time is the average time (over S) in seconds that took each trial to converge. S corresponds to the total trials that converged. n refers to the total number of individuals per region.

Table 6: Simulation Results for Discrete Spatial Heterogeneity: β_{1c}

		C = 49					**C = 100**					**C = 196**				
$\boldsymbol{\theta}_q$	n	**Bias**	**SD**	**RMSE**	**Coverage**	**SE**	**Bias**	**SD**	**RMSE**	**Coverage**	**SE**	**Bias**	**SD**	**RMSE**	**Coverage**	**SE**
	10	-0.516	1.024	1.141	0.898	0.711	-0.538	0.633	0.829	0.959	0.529	-0.367	0.305	0.476	0.828	0.272
	25	-0.092	0.182	0.203	0.950	0.187	-0.069	0.106	0.126	0.970	0.138	-0.074	0.090	0.116	0.910	0.091
$\beta_{1,q=1} = $ -1	50	-0.033	0.103	0.107	0.960	0.110	-0.011	0.069	0.069	1.000	0.081	-0.009	0.050	0.051	0.980	0.053
	80	-0.014	0.068	0.069	0.990	0.081	-0.005	0.065	0.065	0.934	0.061	-0.003	0.033	0.033	1.000	0.040
	100	0.006	0.068	0.068	0.929	0.071	-0.008	0.058	0.059	0.910	0.054	0.004	0.033	0.033	0.969	0.035
	10	-0.039	0.283	0.284	0.761	0.150	-0.055	0.146	0.155	0.837	0.092	-0.065	0.172	0.184	0.828	0.074
	25	-0.011	0.094	0.094	0.900	0.077	-0.006	0.059	0.059	0.950	0.050	-0.010	0.039	0.040	0.940	0.038
$\beta_{1,q=2} = 0$	50	0.000	0.049	0.049	0.960	0.048	0.002	0.030	0.030	0.960	0.032	0.054	0.162	0.171	0.870	0.025
	80	-0.002	0.040	0.040	0.910	0.037	0.056	0.164	0.172	0.882	0.026	-0.001	0.021	0.021	0.930	0.019
	100	0.080	0.212	0.225	0.762	0.034	0.003	0.022	0.022	0.940	0.022	-0.001	0.019	0.019	0.927	0.016
	10	-0.050	0.713	0.711	0.864	0.243	0.018	0.207	0.206	0.959	0.133	0.020	0.119	0.120	0.919	0.081
	25	0.025	0.108	0.111	0.940	0.091	0.025	0.068	0.072	0.960	0.066	0.013	0.049	0.051	0.920	0.045
$\beta_{1,q=3}= 0.5$	50	0.001	0.058	0.058	0.970	0.061	0.012	0.039	0.040	0.950	0.043	-0.055	0.159	0.168	0.860	0.029
	80	0.005	0.048	0.048	0.950	0.047	-0.059	0.159	0.168	0.829	0.032	0.004	0.022	0.023	0.940	0.023
	100	-0.018	0.184	0.184	0.786	0.041	0.004	0.031	0.031	0.910	0.029	0.002	0.018	0.018	0.979	0.021
	10	0.058	0.220	0.226	0.989	0.224	0.009	0.153	0.152	0.929	0.151	0.046	0.115	0.123	0.960	0.105
	25	0.018	0.119	0.120	0.990	0.127	-0.002	0.090	0.089	0.960	0.088	0.015	0.057	0.059	0.950	0.062
$\beta_{1,q=4} = 1$	50	0.012	0.092	0.092	0.940	0.087	0.011	0.070	0.070	0.930	0.062	0.008	0.043	0.043	0.970	0.043
	80	0.006	0.069	0.069	0.940	0.069	0.007	0.052	0.052	0.934	0.049	0.007	0.035	0.036	0.910	0.034
	100	0.049	0.174	0.180	0.976	0.084	-0.007	0.043	0.044	0.940	0.043	0.002	0.030	0.030	0.948	0.031

Note: Bias is $\sum_{s=1}^{S}(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_s)/S$. Standard deviation: SD $= \sqrt{\sum_{s=1}^{S}(\widehat{\boldsymbol{\theta}}_s - \bar{\widehat{\boldsymbol{\theta}}})^2/(S-1)}$. Coverage is the proportion of CIs that contain the true parameter. RMSE $=\sqrt{\sum_{s=1}^{S}(\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_0)^2/S}$. SE correspond to the mean over S of the asymptotic standard error. All models correspond to a binary probit model estimated by SML using 100 Halton Draws. S corresponds to the actual trials that converged (See Table 5).

Table 7: Simulation Results for Discrete Spatial Heterogeneity: β_{2c}

		C = 49					**C = 100**					**C = 196**				
$\boldsymbol{\theta}_q$	n	**Bias**	**SD**	**RMSE**	**Coverage**	**SE**	**Bias**	**SD**	**RMSE**	**Coverage**	**SE**	**Bias**	**SD**	**RMSE**	**Coverage**	**SE**
	10	-0.404	1.006	1.079	0.943	0.671	-0.414	0.724	0.831	0.959	0.466	-0.256	0.222	0.338	0.939	0.250
	25	-0.060	0.194	0.202	0.950	0.177	-0.051	0.125	0.134	0.970	0.130	-0.038	0.087	0.095	0.950	0.085
$\beta_{2,q=1} = -1$	50	0.000	0.096	0.096	0.950	0.103	-0.003	0.074	0.074	0.970	0.078	-0.005	0.053	0.053	0.940	0.052
	80	-0.014	0.080	0.081	0.960	0.080	-0.002	0.060	0.060	0.961	0.060	0.000	0.044	0.043	0.920	0.040
	100	-0.007	0.070	0.070	0.917	0.071	-0.001	0.062	0.062	0.930	0.053	-0.004	0.035	0.035	0.948	0.035
	10	-0.037	0.253	0.254	0.807	0.147	-0.014	0.097	0.097	0.918	0.092	-0.021	0.116	0.117	0.939	0.072
	25	0.008	0.092	0.092	0.920	0.079	0.001	0.050	0.049	0.960	0.051	-0.002	0.035	0.035	0.960	0.039
$\beta_{2,q=2} = -0.5$	50	-0.012	0.050	0.051	0.950	0.052	0.001	0.033	0.033	0.950	0.034	0.051	0.160	0.167	0.840	0.026
	80	-0.001	0.040	0.040	0.950	0.040	0.054	0.155	0.164	0.842	0.027	-0.002	0.023	0.023	0.910	0.020
	100	0.085	0.267	0.279	0.821	0.036	0.001	0.024	0.024	0.950	0.024	0.002	0.015	0.015	0.990	0.018
	10	0.047	0.459	0.459	0.818	0.207	0.050	0.169	0.175	0.898	0.119	0.003	0.106	0.106	0.939	0.073
	25	0.011	0.092	0.092	0.940	0.083	0.032	0.123	0.127	0.930	0.059	0.013	0.042	0.043	0.950	0.041
$\beta_{2,q=3} = 0$	50	0.007	0.056	0.056	0.990	0.055	0.011	0.045	0.046	0.900	0.039	-0.047	0.155	0.161	0.860	0.027
	80	0.001	0.036	0.036	0.990	0.043	-0.056	0.162	0.171	0.816	0.030	0.002	0.020	0.020	0.980	0.021
	100	0.000	0.255	0.253	0.821	0.038	0.000	0.025	0.025	0.950	0.027	0.002	0.017	0.017	0.969	0.019
	10	0.087	0.237	0.251	0.955	0.227	0.030	0.154	0.157	0.949	0.153	0.030	0.111	0.114	0.960	0.107
	25	0.019	0.120	0.121	0.960	0.128	0.012	0.089	0.090	0.960	0.089	0.013	0.071	0.072	0.940	0.062
$\beta_{2,q=4} = 1$	50	0.008	0.088	0.088	0.950	0.087	0.003	0.065	0.065	0.950	0.061	0.002	0.044	0.044	0.950	0.043
	80	0.001	0.072	0.072	0.950	0.069	0.007	0.047	0.047	0.961	0.049	0.006	0.033	0.033	0.970	0.034
	100	0.049	0.161	0.167	0.952	0.086	0.000	0.041	0.040	0.960	0.043	0.001	0.032	0.032	0.917	0.031

Note: Bias is $\sum_{s=1}^{S}(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_s)/S$. Standard deviation: SD $= \sqrt{\sum_{s=1}^{S}(\widehat{\boldsymbol{\theta}}_s - \bar{\widehat{\boldsymbol{\theta}}})^2/(S-1)}$. Coverage is the proportion of CIs that contain the true parameter. RMSE $=\sqrt{\sum_{s=1}^{S}(\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_0)^2/S}$. SE correspond to the mean over S of the asymptotic standard error. All models correspond to a binary probit model estimated by SML using 100 Halton Draws. S corresponds to the actual trials that converged (See Table 5).