THREE ESSAYS ON CONTINUOUS AND DISCRETE SPATIAL HETEROGENEITY

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Mauricio Alejandro Sarrias Jeraldo

May 2016

THREE ESSAYS ON CONTINUOUS AND DISCRETE SPATIAL HETEROGENEITY

Mauricio Alejandro Sarrias Jeraldo, Ph.D.

Cornell University 2016

Continuous and discrete unobserved heterogeneity have been widely used in modeling discrete choice models. In this dissertation I investigate how these modeling strategies can be used to capture and model spatial heterogeneity or locally varying coefficients for different latent structures.

In the first chapter, I outline the main advantages and disadvantages of both continuous and discrete spatial modeling strategies. Then I 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. The results show that the data requirement to achieve lower bias in the continuous case is substantial compared with the discrete case. I have 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.

In the second chapter, I show the Rchoice package for R that allows estimating models with individual heterogeneity for both cross-sectional and panel data. In particular, the package allows binary, ordinal and count response, as well as continuous and discrete covariates. This chapter is a general description of Rchoice and all functionalities are illustrated using real databases.

The last chapter shows how continuous and discrete spatial heterogeneity models can be applied in order to analyze whether monetary subjective well-being evaluations vary across space using a cross-sectional dataset from Chile. The results show that focusing just on the average estimates of compensating variations veils useful local variation. Moreover, the discrete approach shows some weak superiority over the continuous case in terms of model fit and interpretation.

BIOGRAPHICAL SKETCH

Mauricio Sarrias was born in 1986, Antofagasta Chile. He attended Kindergarten and first year of Primary School in a small town called Tocopilla in the North of Chile. He then moved to Antofagasta, a city also located in the north of Chile, at the age of seven, and completed his primary and secondary education at Espana D-59 School, Chile. He finished middle school at Instituto Superior de Comercio at age of 17. He obtained his B.A degree in Business Management in 2008 at Universidad Católica del Norte, Antofagasta, Chile. In 2009, Mr. Sarrias graduated from the same university with a Master of Science degree in Regional Science under the supervision of Professor Patricio Aroca. After obtaining this Master Degree, he was instructor for undergraduate level course for Department of Economic at Universidad Católica del Norte. He also was research assistant for the Institute of Applied Regional Economics at Universidad Católica del Norte. Prior to beginning the PhD program at Cornell University, he was a research assistant for the Development Research Group at the World Bank, Washington D.C., USA. In this position, Mr. Sarrias carried out econometric analyses and data management for the Lead Economist William F. Maloney. From these works he developed an interest on discrete choice models with random parameters. While completing his PhD dissertation under the supervision of Professor Kieran Donaghy, he used some ideas of random coefficient models in order to propose approaches to model spatial heterogeneity. Mauricio is now serving as professor to the Department of Economics at the Universidad Catolica del Norte. Outside of academic, Mr. Mauricio enjoys reading and playing guitar.

To my wife, family and friends.

ACKNOWLEDGEMENTS

I am grateful for my dissertation committee for their patience and insights. I thank Professor Kieran Donaghy, the chairman of my committee, for his supervision throughout the writing of my dissertation. To Professor Ricardo Daziano, member of my committee, I am infinitely grateful for his support and guidance in this process. I owe many thanks to Professor Nancy Brooks for helping me see through the third paper of my dissertation and for important comments and input to my work. I also want to thank you for letting my defense be an enjoyable moment, and for your brilliant comments and suggestions, thanks to you.

In addition to my committee members, I have benefited from many other mentors I had previously, namely, William F. Maloney, Marcelo Lufin, Patricio Aroca, Miguel Atienza and Gianfranco Piras. I thank them all for the continuous support of my Ph.D. study and related research, motivation and immense knowledge.

Also I would like to thank my family for all their love and encouragement. Words cannot express how grateful I am to my beloved wife Yohanna for her patience and unrelenting support during these years. Thank you for your unconditional love. I wish to thank my friends from Cornell. In particular, special thanks are due to Alberto Hernandez and Jazmin Zatarain. Thanks for being there and for making these thousands of kilometers of distance unnoticeable. Thank you!

Finally, I would like to thank my sponsor Becas Chile, from Ministerio de Educación, Government of Chile, because without their financial support this Ph.D. would have never happened.

TABLE OF CONTENTS

Biographical Sketch . . . iii
Dedication . . . iv
Acknowledgements . . . v
Table of Contents . . . vi
List of Tables . . . viii
List of Figures . . . ix

1 Continuous and Discrete Spatial Heterogeneity: Modeling Strategies and Simulation. 1
1.1 Introduction and background . . . 1
1.2 Modeling strategies for spatial heterogeneity . . . 7
1.2.1 Continuous spatial heterogeneity . . . 8
1.2.2 Discrete spatial heterogeneity . . . 16
1.3 Likelihood-based estimation . . . 18
1.4 Region-specific estimates . . . 21
1.5 Simulation experiments . . . 23
1.5.1 Experiment 1: Continuous case . . . 24
1.5.2 Experiment 2: Discrete case . . . 25
1.6 Results . . . 26
1.6.1 Experiment 1: Continuous case . . . 26
1.6.2 Experiment 2: Discrete case . . . 32
1.7 Discussion and conclusion . . . 39

2 Discrete Choice Models with Random Parameters in R: The Rchoice Package 43
2.1 Introduction . . . 43
2.2 Methodology . . . 47
2.2.1 Models with random parameters . . . 47
2.2.2 Extensions: observed heterogeneity and correlation . . . 49
2.2.3 Estimation . . . 51
2.3 An overview of Rchoice package . . . 53
2.4 Drawing from densities . . . 55
2.5 Applications using Rchoice . . . 61
2.5.1 Standard models . . . 61
2.5.2 Random parameters models with cross sectional data . . . 69
2.5.3 Random parameters models with panel data . . . 80
2.5.4 Random parameter model with observed heterogeneity . . . 84
2.5.5 Plotting conditional means . . . 87
2.6 Computational issues . . . 91
2.7 Conclusions . . . 92

3 Do Monetary Subjective Well-Being Evaluations vary Across Space? Comparing Continuous and Discrete Spatial Heterogeneity **94**
3.1 Introduction 94
3.2 Model, assumptions and spatial heterogeneity 98
3.2.1 Model and assumptions 98
3.2.2 Spatial heterogeneity 103
3.3 Econometric modeling 105
3.3.1 Continuous and discrete spatial heterogeneity 106
3.3.2 Estimation 110
3.3.3 Location-specific compensating variations 112
3.4 Data and results 114
3.4.1 Data 115
3.4.2 Fixed coefficients 118
3.4.3 Continuous spatial heterogeneity 121
3.4.4 Discrete spatial heterogeneity 125
3.4.5 Spatial heterogeneity in CV 129
3.5 Discussion and conclusion 137

LIST OF TABLES

1.1 Latent PDFs for different models. 8
1.2 CPU time and trails that converged for experiment 1. 27
1.3 Simulation results for continuous spatial heterogeneity: $C = 1024$. 29
1.4 Simulation results for continuous spatial heterogeneity. 30
1.5 CPU time and trials that converged for experiment 2. 34
1.6 Simulation results for discrete spatial heterogeneity: β_{1c}. 36
1.7 Simulation results for discrete spatial heterogeneity: β_{2c}. 37

2.1 Models estimated by **Rchoice**. 54

3.1 Summary statistics. 116
3.2 Continuous spatial heterogeneity. 119
3.3 Discrete spatial heterogeneity. 127

LIST OF FIGURES

1.1 Continuous distributions for the spatially random parameter. 10
1.2 Conditional estimates of β_{2c} under continuous spatial heterogeneity. 31
1.3 Confident intervals for continuous spatial heterogeneity: β_{2c}. 33
1.4 Conditional estimates of β_{1c} under discrete spatial heterogeneity. . 35
1.5 Conditional estimates discrete spatial heterogeneity: β_{1c} and $C = 196$. 38
1.6 Confidence intervals for discrete spatial heterogeneity: β_{1c}. 40

2.1 Halton vs pseudo-random draws. 56
2.2 Distributions for random coefficients. 60
2.3 Kernel density of the individual's conditional mean. 89
2.4 Individual confidence interval for the conditional means. 90

3.1 Compensating variation for two regions. 104
3.2 Distribution of CVs for individual's covariates. 131
3.3 Distribution of CV for individual's health covariates. 133
3.4 Distribution of CV for perceived neighborhood pollution. 134
3.5 Distribution of CV for perceived public security in the neighborhood. 135
3.6 Distribution of CV for variables at the commune-level. 136
3.7 Compensating variation for disability with 95% CI. 138
3.8 Compensating variation for Unemployed with 95% CI. 139

CHAPTER 1

CONTINUOUS AND DISCRETE SPATIAL HETEROGENEITY: MODELING STRATEGIES AND SIMULATION.

1.1 Introduction and background

In the last few decades there 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, spatial heterogeneity can be defined 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 spatially 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

[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 first 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

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 facts, 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, 2003; Brown and Jones, 1985; McMillen, 1996). This method allows the coefficients to be functions 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

[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).

the parameters represent simple linear expansions of the global parameter over space but more complex expansions as nonlinear and quadratic expressions 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 non-stationarity. 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 esti-

[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).

mating 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 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 one 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

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

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 normally 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, I 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. I 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, I assume 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

[5]See Section 1.2 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).

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] I extend 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. I also discuss the similarities between these modeling strategies with models discussed above, as well as issues that arise in each of them.

I 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, I estimate the models using maximum likelihood and simulated maximum likelihood, depending on the nature of $g()$. A binary probit model is used as the true model governing the true data generating process.

The remainder of this chapter is organized as follows. Section 1.2 discusses the continuous and discrete strategies for modeling spatial heterogeneity. The maximum likelihood estimation procedures are presented in Section 1.3. In Sec-

[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).

tion 1.4 I show how regional-specific estimates can be obtained. The simulation experiments are presented in Section 1.5, and the results analyzed in Section 1.6. Finally, Section 1.7 concludes.

1.2 Modeling strategies for spatial heterogeneity

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

$$\begin{aligned} y_{ci}^* &= \mathbf{x}_{ci}'\boldsymbol{\beta}_c + \epsilon_{ci}, \quad c = 1, ..., C, i = 1, ..., n_c \\ \boldsymbol{\beta}_c &\sim g(\boldsymbol{\beta}_c), \end{aligned} \tag{1.1}$$

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 distributed normal, 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 1.1.

The key element in Equation (2.1) 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

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

Table 1.1: Latent PDFs for different models.

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

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, I discuss the case where $g(\boldsymbol{\beta}_c)$ follows a continuous or discrete distribution, and their respective implications.

1.2.1 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 $\boldsymbol{\theta}$.

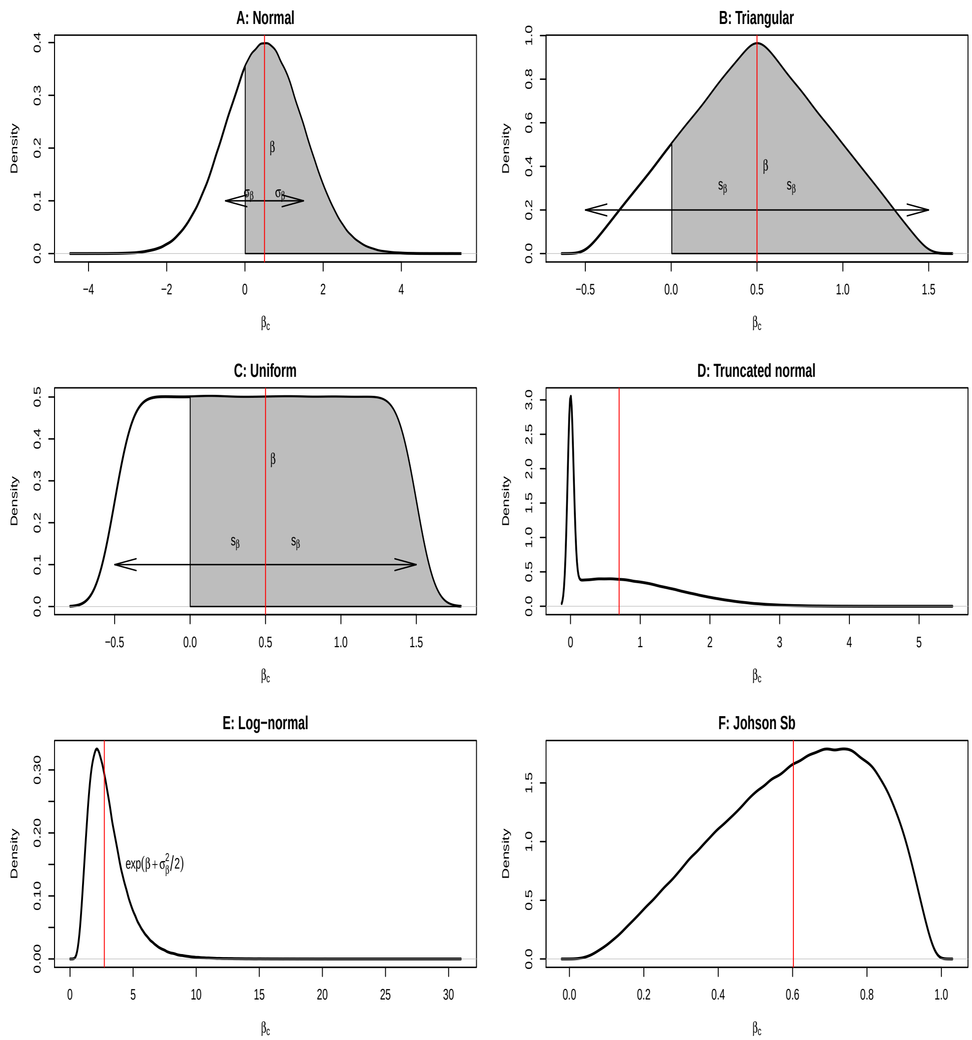
[9] We might allow both intra- and inter-spatial heterogeneity by specifying $\boldsymbol{\beta}_{ci} = \boldsymbol{\beta}_c + \boldsymbol{\delta}_{ci}$, where $\boldsymbol{\beta}_c$ is distributed across regions but not over individuals, and $\boldsymbol{\delta}_{ci}$ is distributed over both individuals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

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, I 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 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 the a priori assumption that there is a proportion of regions with positive coefficients and another proportion with negative ones. As an illustration, Panel A of Figure 1.1 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 shown 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, a large population may suggest agglomeration economies, while in others, 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.

Figure 1.1: Continuous distributions for the spatially random parameter.



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

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 a 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 1.1). 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 1.1. Then the parameter can be constructed as $\beta_c = \beta + s_\beta(2u_c - 1)$ where $u_c \sim \mathrm{U}[0, 1]$ and the parameters $\boldsymbol{\theta} = (\beta, s_\beta)$ are estimated. The new random draw $(2u_c - 1)$ is distributed as $\mathrm{U}[-1, +1]$, 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 and negative coefficients. However, as I 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 $\exp(\beta_c)$, $\exp(\beta_c + \sigma^2_\beta/2)$ and mean $\times \sqrt{\exp(\sigma^2_\beta) - 1}$, 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 1.1.

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 1.1 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

[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).

computed as $\beta_c = \frac{\exp(\beta+\sigma_{\beta}\eta_c)}{1+\exp(\beta+\sigma_{\beta}\eta_c)}$ 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 $\beta+\sigma_{\beta}\eta_c$ 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 the log-normal distribution, but with thinner tails below the 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.

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 $\mathbf{L}\mathbf{L}' = \mathrm{Var}(\boldsymbol{\beta}_c) = \boldsymbol{\Sigma}$. 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 coefficient 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:

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

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:

$$\boldsymbol{\beta}_c = \boldsymbol{\beta} + \boldsymbol{\Pi}\mathbf{z}_c + \mathbf{L}\boldsymbol{\eta}_c \tag{1.2}$$

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 (1.2) can be very useful to consider regions' locations 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 Casetti's spatial expansion method. This approach is used in the simulation exercise.

1.2.2 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 thought of as a classification or segmentation of regions, which are homogeneous in terms of the marginal effects of the variables on the latent process.[11] Formally, the distribution of spatially random parameters is

$$g(\boldsymbol{\beta}_c|\boldsymbol{\theta}_q) = \begin{cases} \boldsymbol{\beta}_1 & \text{with probability } w_{c1} \\ \boldsymbol{\beta}_2 & \text{with probability } w_{c2} \\ \vdots & \vdots \\ \boldsymbol{\beta}_Q & \text{with probability } w_{cQ} \end{cases}, \tag{1.3}$$

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

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 (1.3) 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

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

that the analyst has to choose the distribution of the spatial random parameters a priori, whereas in the discrete heterogeneity case no assumptions are made about the shape of the spatial heterogeneity other than the number of classes. As I will discuss in Section 1.3, another main advantage of this model is that it does not require any simulation-based method to estimate the parameters.

1.3 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

$$\Pr(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\beta}_c) = \prod_{i=1}^{n_c} f^*(y_{ic}|\mathbf{x}_{ic}, \boldsymbol{\beta}_c), \tag{1.4}$$

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 (1.4) 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

$$\begin{aligned} P_c(\boldsymbol{\theta}_q) = f(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\theta}_q) &= \sum_{q=1}^{Q} w_{iq} \left[\prod_{i=1}^{N_c} f^*(y_{ic}|\mathbf{x}_{ic}, \boldsymbol{\beta}_c, \boldsymbol{\theta}_q)\right], &(1.5)\\ P_c(\boldsymbol{\theta}) = f(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\theta}) &= \int_{\boldsymbol{\beta}_c} \left[\prod_{i=1}^{N_c} f^*(y_{ic}|\mathbf{x}_{ic}, \boldsymbol{\beta}_c, \boldsymbol{\theta})\right] g(\boldsymbol{\beta}_c) d\boldsymbol{\beta}_c. &(1.6) \end{aligned}$$

In general, the model with discrete spatial heterogeneity and unconditional pdf given in (1.5) can be easily estimated using maximum likelihood or Expectation-Maximization (EM) algorithm (see for example Ruud, 1991). In the simulation experiment, I 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 (1.6) 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 method 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 (1.6) is calculated. This process is repeated for many draws, and the average over the draws is the simulated probability. Formally, the simulated probability for region c is

$$\widetilde{P}_c(\boldsymbol{\theta}) = \frac{1}{R}\sum_{r=1}^{R}\prod_{i=1}^{N_c}\widetilde{P}_{icr}(\boldsymbol{\theta}), \tag{1.7}$$

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

[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).

function is:

$$\log L_s = \sum_{c=1}^{C} \log \left[\frac{1}{R} \sum_{r=1}^{R} \prod_{i=1}^{N_c} \widetilde{P}_{icr}(\boldsymbol{\theta}) \right]. \tag{1.8}$$

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 (1.7) is 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). As in 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 individuals 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 of 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 a 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 borne in mind when applying these methods.

1.4 Region-specific estimates

In the applied literature it is very common to map 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:

$$f(\boldsymbol{\beta}_c|\mathbf{y}_c, \mathbf{X}_c, \boldsymbol{\theta}) = \frac{f(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\beta}_c)g(\boldsymbol{\beta}_c|\boldsymbol{\theta})}{f(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\theta})} = \frac{f(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\beta}_c)g(\boldsymbol{\beta}_c|\boldsymbol{\theta})}{\int_{\boldsymbol{\beta}_c} f(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\beta}_c)g(\boldsymbol{\beta}_c|\boldsymbol{\theta})d\boldsymbol{\beta}_c}, \tag{1.9}$$

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

$$\bar{\boldsymbol{\beta}}_c = \mathbb{E}\left[\boldsymbol{\beta}_c|\mathbf{y}_c, \mathbf{X}_c, \boldsymbol{\theta}\right] = \frac{\int_{\boldsymbol{\beta}_c} \boldsymbol{\beta}_c f(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\beta}_c) g(\boldsymbol{\beta}_c|\boldsymbol{\theta}) d\boldsymbol{\beta}_c}{\int_{\boldsymbol{\beta}_c} f(\mathbf{y}_c|\mathbf{X}_c, \boldsymbol{\beta}_c) g(\boldsymbol{\beta}_c|\boldsymbol{\theta}) d\boldsymbol{\beta}_c} \tag{1.10}$$

This expectation gives us the conditional mean of the distribution of the spatially random parameter. The simulators of (1.10) for the continuous and discrete case are, respectively:

$$\widehat{\bar{\boldsymbol{\beta}}}_c = \widehat{\mathbb{E}}\left[\boldsymbol{\beta}_c|\mathbf{y}_c, \mathbf{X}_c, \widehat{\boldsymbol{\theta}}\right] = \frac{\frac{1}{R}\sum_{r=1}^{R} \widehat{\boldsymbol{\beta}}_{cr} \prod_{i=1}^{n_c} f^*(y_{ci}|\mathbf{x}_{ci}, \widehat{\boldsymbol{\beta}}_{cr})}{\frac{1}{R}\sum_{r=1}^{R} \prod_{i=1}^{n_c} f^*(y_{ci}|\mathbf{x}_{ci}, \widehat{\boldsymbol{\beta}}_{cr})} \tag{1.11}$$

$$\widehat{\bar{\boldsymbol{\beta}}}_c = \widehat{\mathbb{E}}\left[\boldsymbol{\beta}_c|\mathbf{y}_c, \mathbf{X}_c, \widehat{\boldsymbol{\theta}}_q\right] = \frac{\sum_{q=1}^{Q} \widehat{\boldsymbol{\beta}}_q \widehat{w}_{cq} \prod_{i=1}^{n_c} f^*(y_{ci}|\mathbf{x}_{ci}, \widehat{\boldsymbol{\beta}}_q)}{\sum_{q=1}^{Q} \widehat{w}_{cq} \prod_{i=1}^{n_c} f^*(y_{ci}|\mathbf{x}_{ci}, \widehat{\boldsymbol{\beta}}_q)} \tag{1.12}$$

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):

$$\widehat{V}_c = \widehat{\mathbb{E}}\left[\boldsymbol{\beta}_c^2|\mathbf{y}_c, \mathbf{X}_c, \widehat{\boldsymbol{\theta}}\right] - \left(\widehat{\mathbb{E}}\left[\boldsymbol{\beta}_c|\mathbf{y}_c, \mathbf{X}_c, \widehat{\boldsymbol{\theta}}\right]\right)^2 \tag{1.13}$$

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).

1.5 Simulation experiments

I use two simulation studies to assess the accuracy of the regression coefficients from models with continuous and discrete spatial heterogeneity. To do so, I 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, I 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, I 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 I create databases with $C = \{49, 100, 196\}$ regions. Those numbers are chosen such that regions are equally spaced on a square grid of $\sqrt{C} \times \sqrt{C}$. A similar approach is used for example by Wheeler and Calder (2007) and Páez (2005).
- As I discussed in Section 1.3, it is expected that, if n_c rises without bound, then the conditional means of the parameters, $\widehat{\bar{\boldsymbol{\beta}}}_c$, converge to the true $\boldsymbol{\beta}_c$. Then the question becomes: Given a number of regions, approximately how many individuals per region do we need in order to get conditional means closer to the true $\boldsymbol{\beta}_c$? In order to address this question, I create databases with $n = \{10, 25, 50, 80, 100\}$ individuals in each regions. I limit the experiments to balance the 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. I expand on their work by using random coefficients

where means vary across regions according to their geographical location.

- As I mentioned in Section 1.3, the estimation procedure posits some limitations on the number of individuals per regions due to numerical problems. Thus, I also assess how well the estimation procedures work as the number of individual per region increases.

Given this setting, I have $3 \times 5 = 15$ scenarios in each simulation study. For each scenario I generate $S = 100$ artificial databases (trials). Thus, I 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 held fixed in each scenario. This will allow us to assess whether the conditional density of the regional-specific parameters converges to the true population distribution as n_c becomes bigger.[14]

1.5.1 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)

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

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]

1.5.2 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

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

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

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

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

- 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.

1.6 Results

1.6.1 Experiment 1: Continuous case

I start analyzing the results for the continuous spatial heterogeneity experiment. Table 1.2 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, I note that CPU times increase approximately linearly 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 $n = 80$ I had to discard 64% of the samples because convergence did not occur or because the SML estimates were 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, I only report the results for $n = 10, 25$ and 50.

Table 1.2: 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 1.4 shows the main statistics for the experiment, which correspond to averages over the S trials. I do not report the statistics for the fixed coefficients. Looking at the estimates, I 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

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

when the number of regions increases. Regarding the geographical-coordinates parameters, I 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 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 1.3, 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, I have conducted the same experiment as in Table 1.4, 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 1.3. The general pattern that emerges is that the statistics improve compared with those in Table 1.4. Except for π_{3u}, all the estimates show very low bias and lower variability compared 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, I look into the regional-specific estimates of β_{2c}. Figure 1.2 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 (1.11) and averaged over the S constructed data sets. As discussed in Section 1.4, if n can increase without bound, then both $\widehat{\bar{\beta}}_c$

Table 1.3: 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.

and $\widehat{\mathrm{Var}(\bar{\bar{\beta}}_c)}$ are consistent estimates of β_c, and $\mathrm{Var}(\beta_c)$, respectively. By looking at the densities one 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

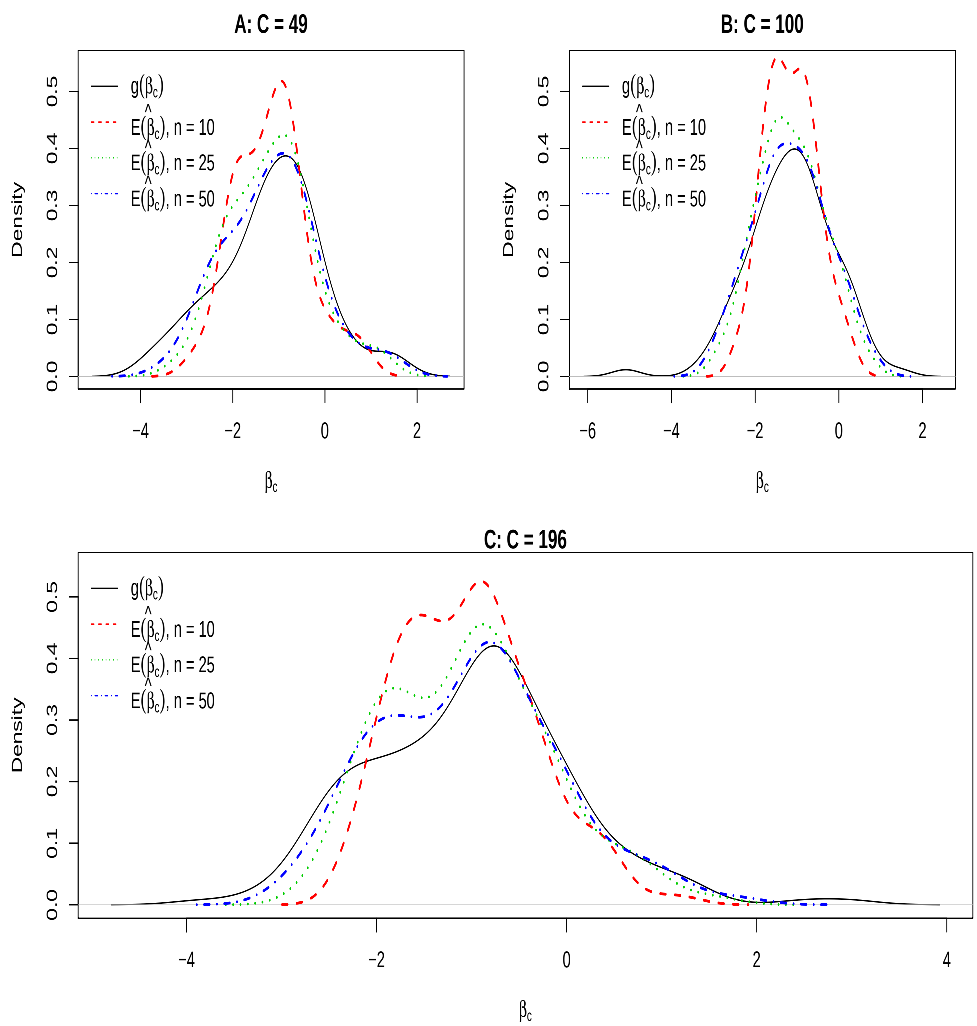
Table 1.4: Simulation results for continuous spatial heterogeneity.

		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 1.2).

better. Another important observation is that, even though there are biases in the model parameters, the conditional estimates are fairly accurate when $n = 50$. I 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 1.2.

Figure 1.2: Conditional estimates of β_{2c} under continuous spatial heterogeneity.



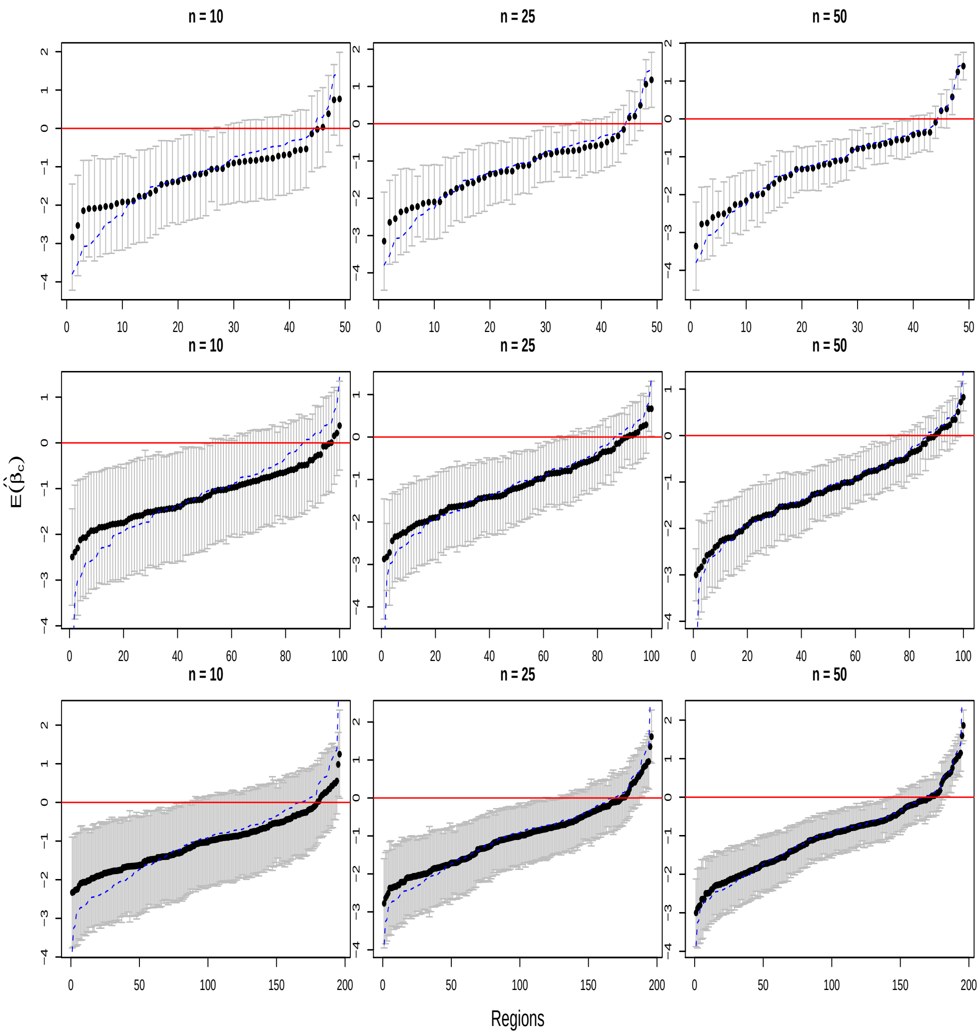
Notes: All the distributions are kernel estimates. The function $g(\boldsymbol{\beta}_c)$ corresponds to the true distribution of the spatially random parameter, which is held fixed during the simulation experiments. Each point in the distribution of the conditional means corresponds to the average over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

I carried out a similar analysis for confidence intervals (CIs) of the regional-specific estimates for β_{2c}. Figure 1.3 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 using Equation (1.13). 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 matters 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 they are more difficult to estimate and then estimator show greater bias.

1.6.2 Experiment 2: Discrete case

Table 1.5 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 in the continuous case.

Figure 1.3: 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.

Results for the discrete spatial heterogeneity experiment are presented in Table 1.6 for β_1 and Table 1.7 for β_2. Given the number of estimation and parameters, it is almost impossible to comment in detail on each coefficient for each class. Given this restriction, I 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

Table 1.5: 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

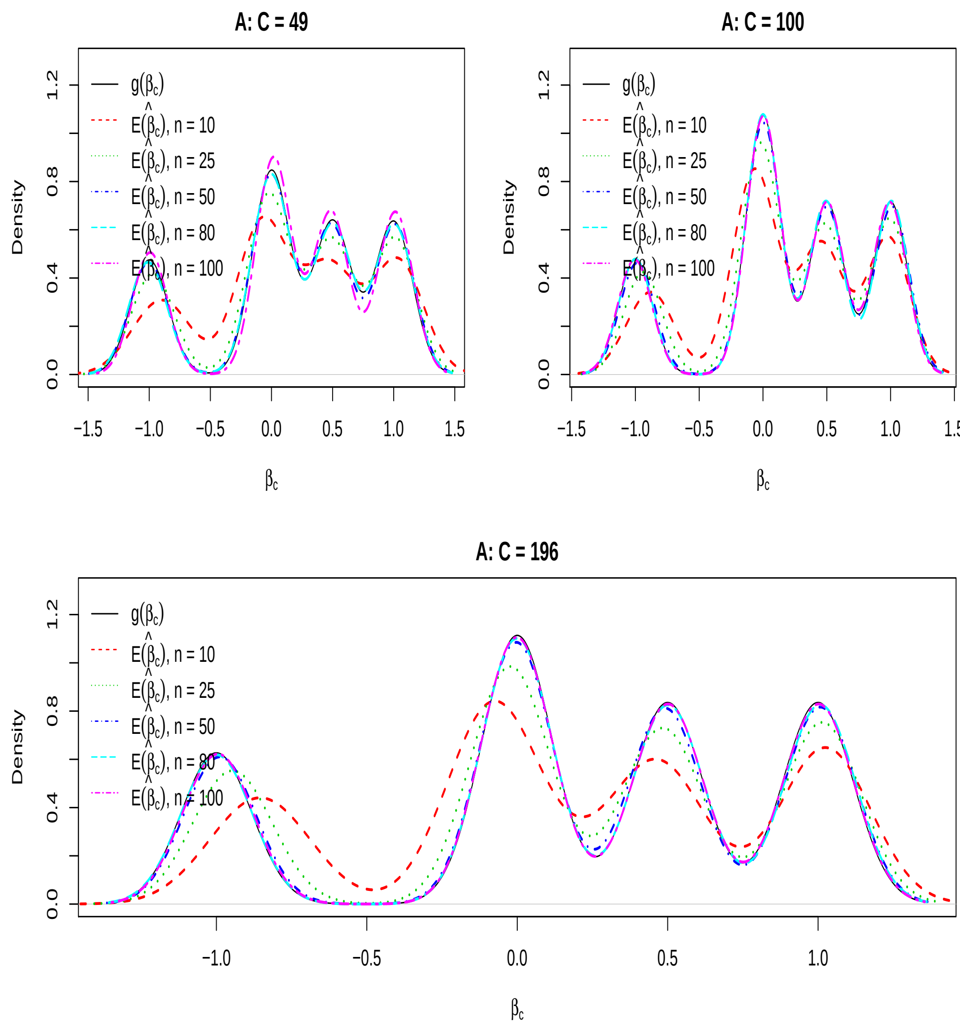
Note: CPU time is the average time (over S) in seconds that took each trial to converge. S corresponds to the total trials that converged. n refers to the total number of individuals per region.

general, the bias is reduced as the number of regions and/or individuals 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 no 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 1.4 displays 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 region. 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,

Figure 1.4: 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 held 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.

and allowing the membership probability to depend on the geographical coordinates is also very convenient to detect clusters of regions, where the clusters are in terms of latent marginal effects. To visualize this, Figure 1.5 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 1.5.2. For instance, the southwest cluster (or class 1)

Table 1.6: 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 1.5).

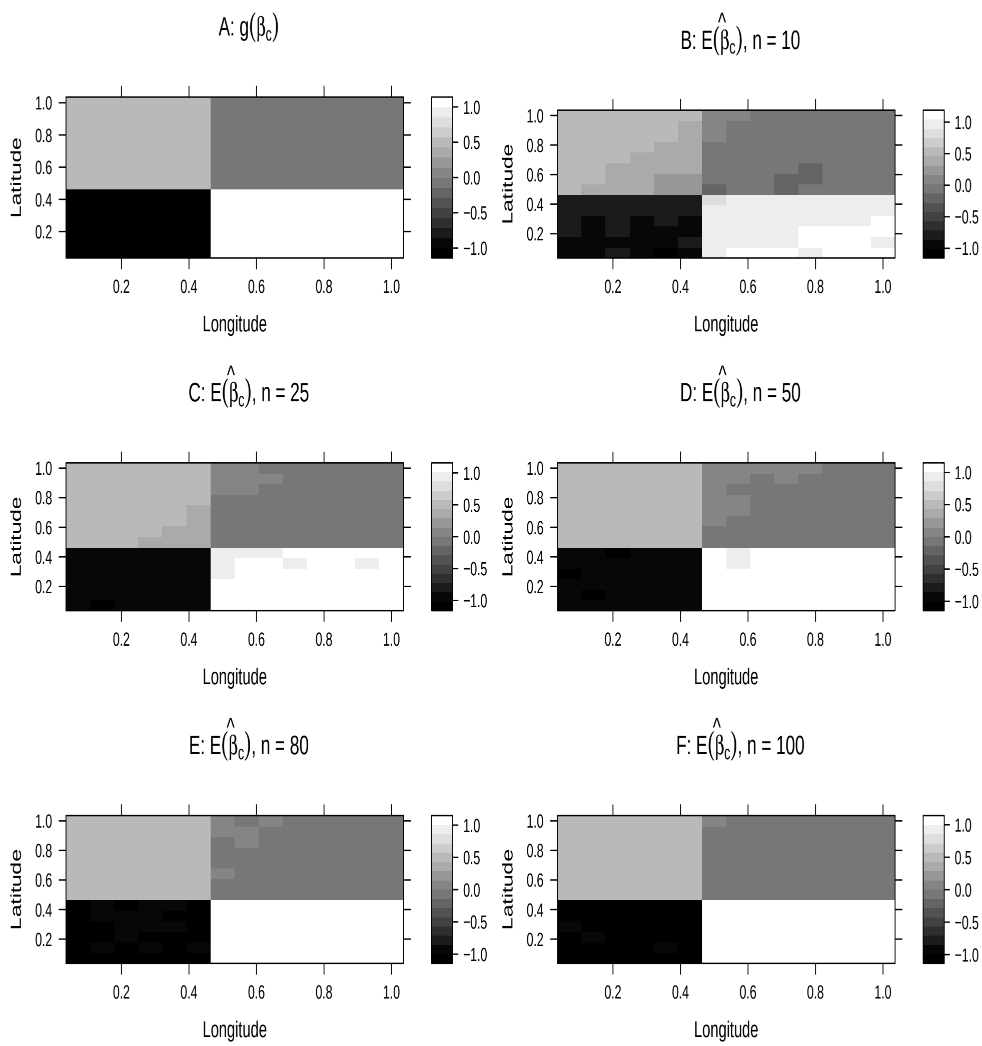
Table 1.7: Simulation results for discrete spatial heterogeneity: β_{2c}.

		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 1.5).

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. I can observe that the spatial pattern of the parameter increasingly resembles the true map surface as the number of individuals per region increases.

Figure 1.5: 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.

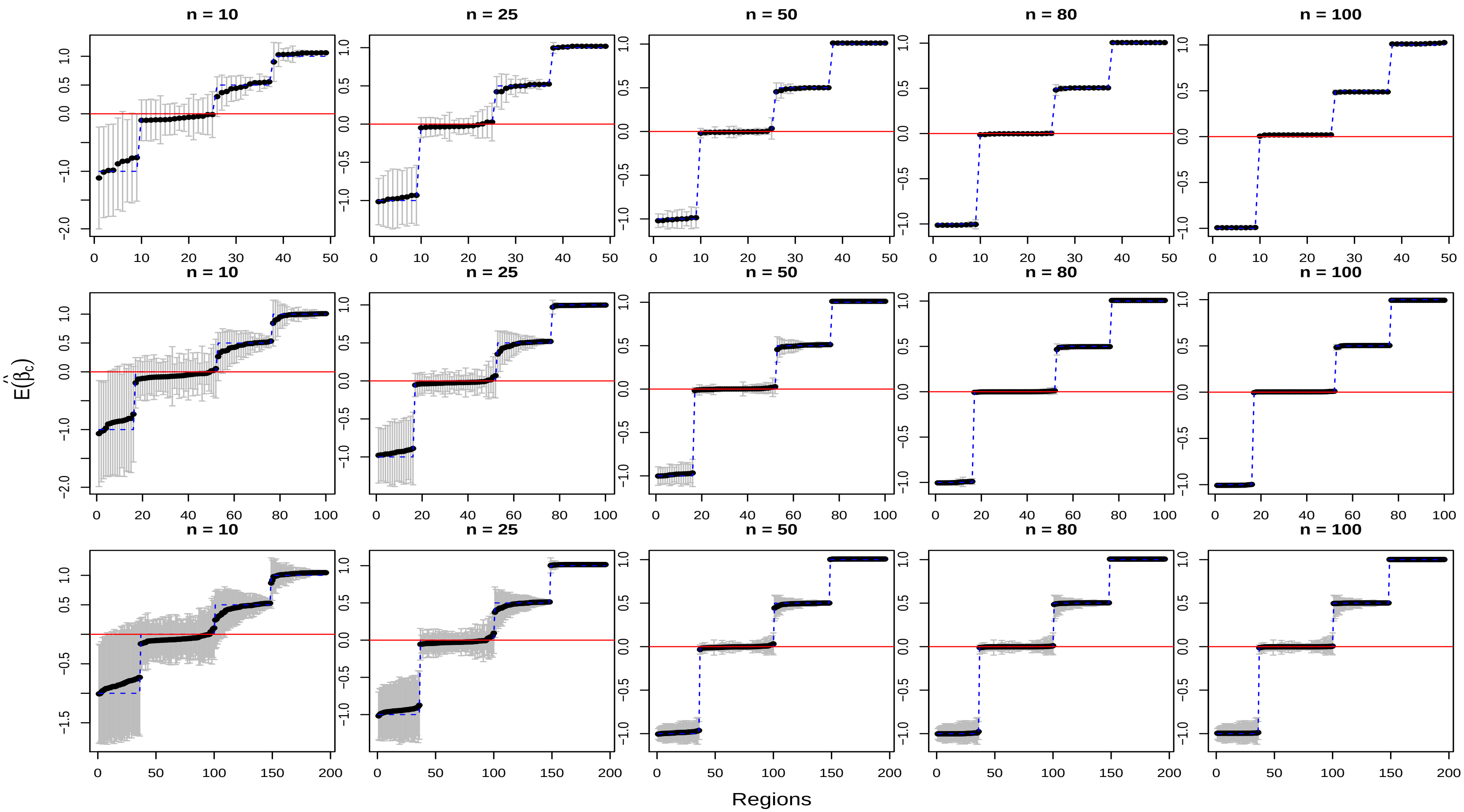
As in the continuous case, I have also plotted the CIs for β_1 under the discrete case. The results are presented in Figure 1.6. Again the first, second and third row present the results for $C = 49$, 100 and 196 respectively; and the dotted-blue line represents the true discrete β_1 for the classes. The CIs are larger when $n = 10$. For the three different values of C, I 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.

1.7 Discussion and conclusion

This chapter contributes to the literature of spatial econometric models that deal with spatially non-stationary processes by examining continuous and discrete unobserved spatial heterogeneity. These two models have been widely used in discrete choice modeling, however I show how these models can be implemented in order to capture and model spatial heterogeneity. One of the main advantages of the two models is that they allow 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. I 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,

Figure 1.6: 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.

thus $g(\boldsymbol{\beta}_c)$ is discrete and modeled in a semi-parametric way. I show how the discrete distribution can be useful to detect clusters 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 does not have a closed form solution when $g(\boldsymbol{\beta}_c)$ is continuous. Therefore, we need to simulate this probability and estimate the parameters using SML, 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 increase linearly with the number of classes.

I 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: the 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, I 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 sample size.

This work can be extended in different ways. First, one of the main concerns and limitations 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 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 for both models are needed in order to understand their strengths and weaknesses for estimating models with locally varying coefficients.

CHAPTER 2

DISCRETE CHOICE MODELS WITH RANDOM PARAMETERS IN R: THE RCHOICE PACKAGE

2.1 Introduction

Discrete choice models or qualitative choice models are intended to describe, explain and predict choices between two or more discrete alternatives, such as buying a car or not, choosing between heating systems, or choosing among different occupations among other applications (Train, 2009). One of the main feature of these models is that the choice made by each individual can be derived under the assumption of utility-maximization behavior.[1] Essentially, the decision maker chooses the alternative that has the higher satisfaction utility given a set of attributes of the person, the attributes of the alternatives, plus a random term intended to capture the factors that affect utility but are not included. For example, the utility that individual i obtains from alternative $j = 1, ..., J$ can be written as $U_{ij} = V_{ij} + \epsilon_{ij}$, where V_{ij} is the deterministic part of the utility that depends upon observable characteristics and ϵ_{ij} is the random part. Individual i chooses the alternative j if $U_{ij} > U_{ik}$ for all $k \neq j$. Once the distribution of ϵ_{ij} and the nature of the observable output decision are specified, a probabilistic model can be used in order to estimate the parameters of the behavioral process and the probability of choosing some alternative. This formulation known as the random utility model (RUM) has been the standard approach in order to derive the conditional Logit Model (McFadden, 1974). However, it can also be applied to describe the relation

[1] As explained by Train (2009), this derivation assures that the model is consistent with utility maximization, but it does not preclude the model from being consistent with other forms of behaviors.

of explanatory variables and the decision when the observed output is binary, ordered or count data. For these types of dependent variables, the traditional binary Logit/Probit model, ordered Probit/Logit, and Poisson model for count data can been applied. For a general overview of these models see for example Long (1997) and Winkelmann and Boes (2006).

One important modeling shortcoming of these methods is the inherent assumption of a fixed and unique coefficient for all individuals in the sample, which might not be realistic given that individuals are intrinsically heterogeneous. As an example, consider the effect of the number of young children on scientists' productivity measured as the number of published articles. Since having more kids may imply fewer hours available to work, we would expect, on average, a negative correlation between the number of children and the level of publications. This negative relationship is global, but it neglects the fact that for some scientists having more kids might imply having a more organized life-style and hence increase their productivity (see for example Krapf et al., 2014). Thus, the negative coefficient might hide significant individual heterogeneity resulting in misleading inferences for a subgroup of individuals with a positive coefficient. Similarly, we might assume that the effect is negative for all scientists, but the magnitude of the detrimental effect might vary across the sample. In this case, there exists individual heterogeneity but only in the negative domain of the coefficient. Finally, we might also find that the coefficient is zero (not significant). In such case, we would conclude that having young children is not important for productivity. Nonetheless, this may be due to the fact that heterogeneity among scientists in the sample cancels out positive and negative effects. Under these circumstances a model that allows for individual heterogeneity may be more appropriate.

Thanks to the emergence of more powerful computers and the development of simulation-based models, much of the recent progress in regression models with limited dependent variables has focused on more realistic behavioral models that allow individual heterogeneity in the parameters. The modeling strategy to accommodate such heterogeneity is assuming that coefficients vary randomly across individuals according to some continuous distribution, denoted by $g(\boldsymbol{\theta})$. Since the distribution is unknown and must be specified a priori by the researcher, this type of individual heterogeneity is usually labeled as 'unobserved heterogeneity' in the literature. All the information of the unobserved heterogeneity is captured by the parameters of the distribution $\boldsymbol{\theta}$, which usually represent the mean and variance of the coefficient. The goal is to estimate those parameters in order to get a profile of individual heterogeneity.

Although the random parameter approach has been widely applied to the Multinomial Logit model (Train, 2009; Hensher and Greene, 2003), there are just a few applications for ordinal (Falco et al., 2015; Greene and Hensher, 2010b) and count models (Gourieroux et al., 1984; Greene, 2007; Anastasopoulos and Mannering, 2009). The scarcity of applications to models other than the Multinomial Logit appears to be driven by the lack of statistical software that enables one to estimate random parameter models for binary, ordinal, and count response. To my knowledge, only the commercial programs LIMDEP (Greene, 2015a) and NLOGIT (Greene, 2015b) are able to estimate those types of models in a concise and flexible manner. The **Rchoice** (Sarrias, 2015b) package for R (Team, 2015) is intended to make these estimation methods available to the general public and practitioners in a friendly and flexible way.

The aim of this chapter is to present the functionalities of **Rchoice** for estimating ordered, count and binary choice models with random parameters. Its current version 0.3 allows estimating cross-sectional and panel (longitudinal) data models and includes also new functionalities to obtain the standard errors of the variance-covariance matrix of the random parameters. All models in **Rchoice** are estimated using simulated maximum likelihood (SML), which is very flexible for estimating models with a large number of random parameters (Train, 2009). An additional characteristic of **Rchoice** is the ability to retrieve the individual conditional estimates of either the random parameters or compensating variations.

Rchoice is also intended to complement other related packages in R. For example, there exist several packages to estimate binary, count and ordered models with fixed parameters. The `glm` function (Team, 2015) allows one to estimate different kinds of discrete choice models such as Poisson and binary models. The function `probit` from the **micEcon** (Henningsen, 2014) package allows one to estimate Probit model. Moreover, the function `polr` from the package **MASS** (Venables and Ripley, 2002) allows one to estimate ordered Probit and Logit models. The advantage of **Rchoice** is that allows more flexibility in the optimization routines, which might improve the convergence speed. Furthermore, **Rchoice** offers an alternative approach to fitting random effects models in the context of panel data to those already programmed in **pglm** (Croissant, 2013) and **lme4** (Doran et al., 2007). Regarding random parameter models, **mlogit** (Croissant et al., 2012), **RSGHB** (Dumont et al., 2014) and **gmnl** (Sarrias and Daziano, 2015) allow estimating models with individual heterogeneity in the context of Multinomial Logit model in a very similar fashion to **Rchoice**. Other packages such as **FlexMix** (Grün and Leisch, 2008) and **PReMiuM** (Liverani et al., 2015) allow one to estimate models with individual heterogeneity by assuming that $g(\boldsymbol{\theta})$ is discrete or a mixture of distributions. All these packages available in R cover almost all possibilities of estimating discrete choice models with random parameters or individual heterogeneity. The **Rchoice** is available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=Rchoice`.

This chapter is organized as follows. Section 2.2 briefly explains the methodology for models with random parameters. Section 2.3 gives a general description of the functions in **Rchoice**. Section 2.4 shows how **Rchoice** handles the random coefficients. All functionalities of **Rchoice** using real databases are presented in Section 2.5. Section 2.6 explains some computational issues that may arise when estimating random parameter models using SML. Finally, Section 2.7 concludes.

2.2 Methodology

2.2.1 Models with random parameters

In order to develop and motivate the idea behind random parameter models, consider the following latent process

$$
\begin{aligned}
y_{it}^* &= \mathbf{x}_{it}'\boldsymbol{\beta}_i + \epsilon_{it}, \quad i = 1, ..., n; t = 1, ..., T_i \\
\boldsymbol{\beta}_i &\sim g(\boldsymbol{\beta}_i|\boldsymbol{\theta}),
\end{aligned}
\tag{2.1}
$$

where y_{it}^* is a latent (unobserved) process for individual i in period t, $\mathbf{x}_{it}$ is a vector of covariates, and ϵ_{it} is the error term. Note that the conditional probability density function (pdf) of the latent process $f(y_{it}^*|\mathbf{x}_{it}, \boldsymbol{\beta}_i)$ is determined once the nature of the observed y_{it} and the population pdf of ϵ_{it} is known: if y_{it} is binary and ϵ_{it} is distributed as normal, then the latent process becomes the traditional Probit model; if y_{it} is an ordered categorical variable and ϵ_{it} is logistically distributed, then the traditional ordered Logit model arises. Formally, the pdf for binary, ordered, and Poisson model are, respectively

$$
f(y_{it}^*|\mathbf{x}_{it}, \boldsymbol{\beta}_i) = \begin{cases} \left[F\left(\mathbf{x}_{it}'\boldsymbol{\beta}_i\right)\right]^{y_{it}}\left[1 - F\left(\mathbf{x}_{it}'\boldsymbol{\beta}_i\right)\right]^{1-y_{it}} \\ \prod_{j=1}^{J}\left[F(k_j - \mathbf{x}_{it}'\boldsymbol{\beta}_i) - F(k_{j-1} - \mathbf{x}_{it}'\boldsymbol{\beta}_i)\right]^{y_{itj}} \\ \frac{1}{y_{it}!}\exp\left[-\exp\left(\mathbf{x}_{it}'\boldsymbol{\beta}_i\right)\right]\exp\left(\mathbf{x}_{it}'\boldsymbol{\beta}_i\right)^{y_{it}} \end{cases}. \tag{2.2}
$$

For the binary and ordered models, $F(\cdot)$ represents the cumulative distribution function (cdf) of the error term, which $F(\epsilon) = \Phi(\epsilon)$ for Probit and $F(\epsilon) = \Lambda(\epsilon)$ for Logit.[2] For the ordered model, κ_j represents the threshold for alternative $j = 1, ..., J-1$, such that $\kappa_0 = -\infty$ and $\kappa_J = +\infty$.

In the structural model given by Equation 2.1, we allow the vector coefficient $\boldsymbol{\beta}_i$ to be different for each individual in the population. In other words, the marginal effect on the latent dependent variable is individual-specific. Nevertheless, we do not know how these parameters vary across individuals. All we know is that they vary according to the population pdf $g(\boldsymbol{\beta}_i|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents the moments of the distribution such as the mean and the variance, which must be estimated. A fully parametric model arises once $g(\boldsymbol{\beta}_i|\boldsymbol{\theta})$ and the distribution of ϵ are specified. In Section 2.4 I detail the distributions allowed by **Rchoice**.

For simplicity in notation, assume that the coefficient vector is independently normally distributed, so that $\beta_k \sim N(\beta_k, \sigma_k^2)$ for the kth element in $\boldsymbol{\beta}_i$. Note that each coefficient can be written as $\beta_{ki} = \beta_k + \sigma_k\omega_i$, where $\omega_i \sim N(0, 1)$, or in vector form as $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{L}\boldsymbol{\omega}_i$, where $\mathbf{L}$ is a diagonal matrix that contains the standard deviation parameters, σ_k. All the information about the individual heterogeneity

[2]For the Logit model $\Lambda(\epsilon) = \exp(\epsilon)/\left[1 + \exp(\epsilon)\right]$.

for each individual attribute is captured by the standard deviation parameter σ_k. If $\sigma_k = 0$, then the model is reduced to the fixed parameter model, but if it is indeed significant then it would reveal that the relationship between x_{itk} and y_{it} is heterogeneous and focusing just on the central tendency β_k alone would veil useful information. It is useful to note that the random effect model is a special case in which only the constant is random (see Section 2.5.3).

2.2.2 Extensions: observed heterogeneity and correlation

One important and straightforward extension of the random parameter model is to allow the coefficients to be correlated. In this case, $\mathbf{L}$ is a lower triangular matrix (also known as the Cholesky matrix), which produces the covariance matrix of the random parameters, $\mathbf{L}\mathbf{L}^\top = \mathbf{\Sigma}$. Furthermore, observed heterogeneity can be also accommodated by allowing the parameter heterogeneity to be partly systematic in terms of observed variables. This type of model is also known as a hierarchical model (Greene and Hensher, 2010a). Formally, the parameter vector can be written as

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{\Pi}\mathbf{s}_i + \mathbf{L}\boldsymbol{\omega}_i, \tag{2.3}$$

where $\mathbf{\Pi}$ is a matrix of parameters, $\mathbf{s}_i$ is a vector of covariates that do not vary across time, and $\boldsymbol{\omega} \sim N(\mathbf{0}, \mathbf{I})$. Then, the mean of the parameters is $\mathbb{E}(\boldsymbol{\beta}_i) = \boldsymbol{\beta} + \mathbf{\Pi}\mathbf{s}_i + \mathbf{L}\mathbb{E}(\boldsymbol{\omega}) = \boldsymbol{\beta} + \mathbf{\Pi}\mathbf{s}_i$ and its covariance is $\mathrm{Var}(\boldsymbol{\beta}_i) = \mathbb{E}(\mathbf{L}\boldsymbol{\omega}(\boldsymbol{\omega}\mathbf{L})^\top) = \mathbf{L}\mathbb{E}(\boldsymbol{\omega}\boldsymbol{\omega}^\top)\mathbf{L}^\top = \mathbf{L}\mathbf{I}\mathbf{L}^\top = \mathbf{L}\mathbf{L}^\top = \mathbf{\Sigma}$.

As an illustration of the formulation above, suppose that we are interested in modeling some latent process, which is explained by two variables. The parameters associated with each observed variable, β_{1i} and β_{2i}, are not fixed, but rather they vary across individuals and are correlated. Furthermore, suppose that the means of the random parameters depend upon some observed individual socio-economic variables: S_i, B_i and C_i. Then, the parameters can be written as:

$$\begin{aligned}
\beta_{1,i} &= \beta_1 + \pi_{1,1}S_i + \pi_{1,2}B_i + \pi_{1,3}C_i + s_{11}\omega_{1,i} \\
\beta_{2,i} &= \beta_2 + \pi_{2,1}S_i + \pi_{2,2}B_i + \pi_{2,3}C_i + s_{21}\omega_{1,i} + s_{22}\omega_{2,i}
\end{aligned}$$

or in vector form:

$$\begin{pmatrix} \beta_{1,i} \\ \beta_{2,i} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \pi_{1,1} & \pi_{1,2} & \pi_{1,3} \\ \pi_{2,1} & \pi_{2,2} & \pi_{2,3} \end{pmatrix} \begin{pmatrix} S_i \\ B_i \\ C_i \end{pmatrix} + \begin{pmatrix} s_{11} & 0 \\ s_{21} & s_{22} \end{pmatrix} \begin{pmatrix} \omega_{1,i} \\ \omega_{2,i} \end{pmatrix}.$$

Since the mean of the random parameter is a function of observed variables, this specification relaxes the assumption of homogeneity across individuals in terms of the means. The unobserved part accounts for all other individual-specific factors that cannot be captured for the observed variables. Note that the variance-covariance matrix of the random parameters is:

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\top = \begin{pmatrix} s_{11} & 0 \\ s_{21} & s_{22} \end{pmatrix} \begin{pmatrix} s_{11} & s_{21} \\ 0 & s_{22} \end{pmatrix} = \begin{pmatrix} s_{11}^2 & s_{11}s_{21} \\ s_{21}s_{11} & s_{21}^2 + s_{22}^2 \end{pmatrix},$$

and the conditional mean vector is $\mathbb{E}(\boldsymbol{\beta}_i|\mathbf{s}_i) = \boldsymbol{\beta} + \mathbf{\Pi}\mathbf{s}_i$, which varies across individuals because of $\mathbf{s}_i$. Illustrations with real data for the model with correlated random parameters and with observed heterogeneity are presented in Section 2.5.2 and Section 2.5.4, respectively.

2.2.3 Estimation

In this section, I explain the simulated maximum likelihood (SML) procedure used by **Rchoice** to estimate models with random parameters. For a more complete treatment of SML see for example Gourieroux and Monfort (1997); Lee (1992); Hajivassiliou and Ruud (1994) or Train (2009).

Let $\mathbf{y}_i = \{y_{i1}, y_{i2}, ..., y_{iT_i}\}$ be the sequence of choices made by individual i. Assuming that individuals are independent across time, the joint pdf, given $\boldsymbol{\beta}_i$, can be written as

$$\Pr(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_i) = \prod_{t=1}^{T_i} f(y_{it}^*|\mathbf{x}_{it}, \boldsymbol{\beta}_i), \tag{2.4}$$

where $f(y_{it}^*|\mathbf{x}_{it}, \boldsymbol{\beta}_i)$ is given in (2.2) for each model. Since $\boldsymbol{\beta}_i$ is unobserved, we need to integrate it out of the joint density. The unconditional pdf will be the weighted average of the conditional probability (2.4) evaluated over all possible values of $\boldsymbol{\beta}$, which depends on the parameters of the distribution of $\boldsymbol{\beta}_i$:

$$P_i(\boldsymbol{\theta}) = \int_{\boldsymbol{\beta}_i} \Pr(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_i) g(\boldsymbol{\beta}_i) d\boldsymbol{\beta}_i. \tag{2.5}$$

The probability in (2.5) has no closed-form solution, that is, it is difficult to integrate out the random parameter and hence it is difficult to perform maximum likelihood (ML) estimation. However, ML estimation may still be possible if we instead use a good approximation $\widetilde{P}_i(\boldsymbol{\theta})$ of $P_i(\boldsymbol{\theta})$ to form a likelihood function.

But, how can we obtain $\widetilde{P}_i(\boldsymbol{\theta})$? A good approximation can be obtained by **Monte Carlo integration**. This procedure provides an alternative to deterministic numerical integration. Here we can '*simulate*' the integration using random draws from the distribution $g(\boldsymbol{\beta}_i)$.[3] For a given value of the parameters $\boldsymbol{\theta}$, a value of $\boldsymbol{\beta}_i$ is drawn from its distribution. Using this draw of $\boldsymbol{\beta}_i$, $P_i(\boldsymbol{\theta})$ from Equation (2.5) is calculated. This process is repeated for many draws, and the average over the draws is the simulated probability. Formally, the simulated probability for individual i is

$$\widetilde{P}_i(\boldsymbol{\theta}) = \frac{1}{R}\sum_{r=1}^{R}\widetilde{\Pr}(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_{ir}), \tag{2.6}$$

where $\widetilde{\Pr}(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_{ir})$ is the simulated probability for individual i in period t evaluated at the rth draw of $\boldsymbol{\beta}_i$, and R is the total number of draws. Given independence over i, the SML estimator is the value $\boldsymbol{\theta}$ that maximizes:

$$\widehat{\boldsymbol{\theta}}_{SML} \equiv \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max} \quad \sum_{i=1}^{N}\log \widetilde{P}_i(\boldsymbol{\theta}).$$

Lee (1992) and Hajivassiliou and Ruud (1994) show that under regularity conditions, the SML 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.

Rchoice uses the **maxLik** package (Henningsen and Toomet, 2011) to perform ML and SML estimation procedures. All models with random parameters are estimated using SML, while models with fixed parameters are estimated by ML. Furthermore, **Rchoice** uses analytical gradients and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (the default) to iteratively solve the SML. For further details about the computation of SML see Section 2.6.

[3] Another numerical approximation is Gauss-Hermite quadrature. However, it has been documented that for models with more than 3 random parameters SML performs better (Train, 2009).

2.3 An overview of Rchoice package

After installation, **Rchoice** is loaded by typing:

```
R> library("Rchoice")
```

The main function in the package is `Rchoice`, which enables us to estimate the models. The arguments for this function are the following:

```
R> args(Rchoice)
function (formula, data, subset, weights, na.action, family,
    start = NULL, ranp = NULL, R = 40, haltons = NA, seed = 10,
    correlation = FALSE, panel = FALSE, index = NULL, mvar = NULL,
    print.init = FALSE, init.ran = 0.1, gradient = TRUE, ...)
NULL
```

The `formula` argument is a symbolic description of the model to be estimated and consists of two parts. The first one is reserved for standard variables with fixed and random parameters. The second part is reserved for variables that enter the mean of the random parameters ($\mathbf{s}_i$ from Equation (2.3)). The usage of the second part is further explained in Section 2.5.4. The models are specified by the argument `family`. The models estimated by `Rchoice` and the corresponding `family` function are presented in Table 2.1.

The `family` for binary and Poisson models comes from **stats** package (Team, 2015), while those for ordered models are included in **Rchoice**.

The main arguments for the control of the random parameters are `ranp`, `R`, `haltons`, `seed`, `correlation` and `mvar`. The argument `ranp` is a vector that

Table 2.1: Models estimated by **Rchoice**.

Model	*Function*
Poisson	`family = poisson`
Binary Probit	`family = binomial("probit")`
Binary Logit	`family = binomial("logit")`
Ordered Probit	`family = ordinal("probit")`
Ordered Logit	`family = ordinal("logit")`

specifies the distribution of the random parameters. The distributions allowed by `Rchoice` are presented in Section 2.4 and illustrated in Section 2.5.2. `R` specifies the number of draws used for the simulation of the probabilities. `haltons` indicates the type of draws used in the simulation procedure and it is further explained in Section 2.4. The argument `seed` sets the seed for the pseudo-random draws, being 10 the default. If `correlation = TRUE`, the correlated random parameter model presented in Section 2.2.2 is estimated. For an example see Section 2.5.2. Finally, the argument `mvar` is a named list that indicates how the variables that modify the mean of the random parameters enter in each random parameter. This feature is illustrated in Section 2.5.4.

The main arguments for panel or longitudinal data bases are `panel` and `index`. If `panel = TRUE`, then the panel structure of the data is taken into account. `index` is a string variable indicating the 'id' for individuals in the data. Models with panel structure are presented in Section 2.5.3.

Finally, the arguments `print.init`, `init.ran` and `gradient` are intended to control the optimization procedure. If `print.init = TRUE`, then the initial values for the optimization procedure are printed. The argument `init.ran` sets the initial values for the standard deviation of the random parameters, $\sigma_k, \forall k$. Numerical instead of analytical gradients are used in the optimization of either ML or SML if `gradient = FALSE`. Other arguments that control the optimization can be passed to **maxLik** package. A more detailed discussion of these arguments are presented in Section 2.6.

2.4 Drawing from densities

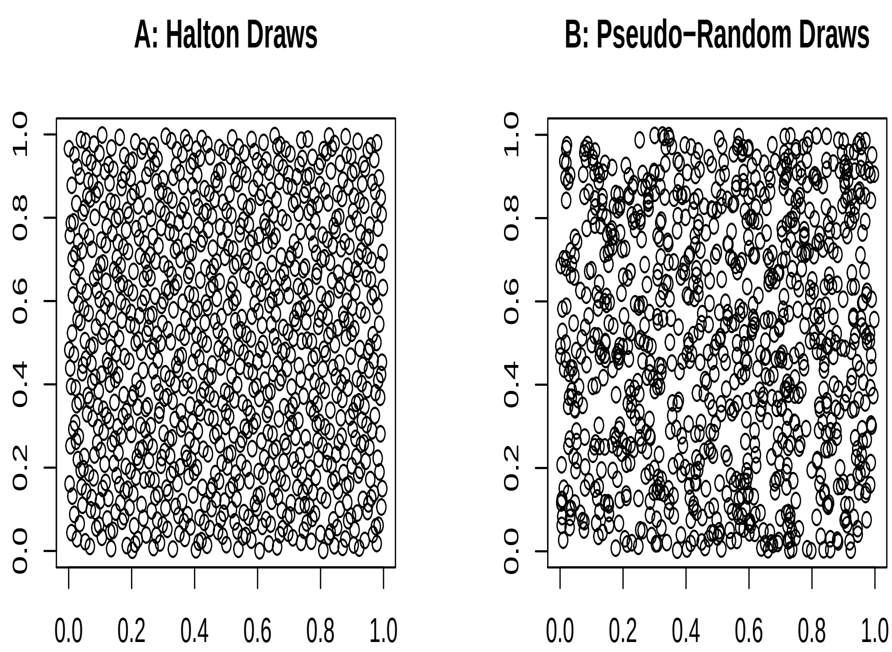
As mentioned in Section 2.2.1, the distribution of the random parameters $g(\boldsymbol{\beta}_i)$ can take any shape, but it must be chosen a priori by the researcher. This is a critical step in applied work. As stated by Hensher and Greene (2003), *"distributions are essentially arbitrary approximations to the real behavioral profile. The researcher chooses a specific distribution because he has a sense that the "empirical truth" is somewhere in their domain"*. Therefore, some prior theoretical knowledge about the expected domain of the coefficients may lead to a more appropriate choice of the distributions. The distributions allowed by **Rchoice** are the following: normal, triangular, uniform, log-normal, truncated normal and Johnson S_b. These distributions are explained in further detail below.

Another important issue is that good performance of SML requires a very large number of draws. The main drawback to this approach is that with large samples and complex models, the maximization of SML can be very time consuming. Researchers have gained speed with no degradation in simulation performance through the use of a smaller number of Halton draws (Bhat, 2001; Train, 2000).[4] The Halton sequence is constructed based on a deterministic method that

[4]There is no consensus about the number of draws that has to be used in applied work. Bhat (2001)'s Monte Carlo analysis found that the precision of the estimated parameters was smaller using 100 Halton draws than 1000 pseudo-random number in the context of Mixed Logit. However, as suggested by Hensher and Greene (2003), the best approach is to estimate models over a range of draws (e.g., 25, 50, 100, 250, 1000, 2000 draws) and analyze the stability and precision of the parameters. In general, and depending in the application, the results should

uses prime numbers as its base (see Train, 2009, for further details). The idea is that, instead of taking independent random draws, simulation can potentially be improved by selecting evaluation points more systematically and with better coverage (Sándor and Train, 2004). Figure 2.1 shows two sequences of 1,000 Halton draws (Panel A) and two sequences of 1,000 pseudo-random draws (Panel B). The Halton draws are based on prime numbers 2 and 3. It can be observed that the Halton draws have better coverage of the unit square than the pseudo-random draws. This characteristic of the Halton sequence ensures a better coverage of the multidimensional area of integration and reduces the computation time of the SML.

Figure 2.1: Halton vs pseudo-random draws.



Rchoice handles Halton or pseudo-random draws in the following way. Suppose

stabilize after about 500 draws.

that there are K random parameters. Then, the K elements of $\boldsymbol{\omega}_{ir}$ are drawn as follows. We begin with a K random vector $\boldsymbol{\omega}_{ir}$, that is:

- K independent draws from the standard uniform $(0, 1)$ distribution or,
- K independent draws from the mth Halton sequence, where m is the mth prime number in the sequence of K prime numbers.

An important attribute of the Halton values is that they are also distributed in the $(0, 1)$ interval as shown in Figure 2.1. Then, the primitive draw (pseudo or Halton draws) is transformed to the distribution specified by the user as follows:

- $u_{k,ir} \sim U(0, 1)$: primitive draw from Halton or pseudo-random number generator.
- $w_{k,ir} = \Phi^{-1}(u_{k,ir}) \sim N(0, 1)$.

Using these two primitive draws, **Rchoice** creates the random parameters as follows:

1. Normal distribution:

$$\begin{aligned} \beta_{k,ir} &= \beta_k + \sigma_k w_{k,ir} \\ w_{k,ir} &\sim N(0, 1) \end{aligned}$$

   where β_k and σ_k are estimated. Then, $\beta_{k,i} \sim N(\beta_k, \sigma_k^2)$. Since the domain of the normal distribution is $(-\infty, +\infty)$, assuming a given coefficient to follow a normal distribution is equivalent to making an a priori assumption that there is a proportion of individuals with a positive coefficient and another proportion with negative ones (see Panel A of Figure 2.2). For example, the proportion of positive coefficients can be computed as $\Phi(\widehat{\beta}_k/\widehat{\sigma}_k)$. The main disadvantage of the normal distribution is that it has infinite tails, which may result in some individuals having implausible extreme coefficients. If this is the case, the triangular or uniform distribution may be more appropriate.

2. Triangular distribution:

$$\begin{aligned}\beta_{k,ir} &= \beta_k + \sigma_k v_{k,ir}\\ v_{k,ir} &\sim \mathbb{1}(u_{k,ir} < 0.5)\left(\sqrt{2u_{k,ir}} - 1\right) + \mathbb{1}(u_{k,ir} \geq 0.5)\left[1 - \sqrt{2(1-u_{k,ir})}\right]\end{aligned}$$

where $\mathbb{1}(\cdot)$ is an indicator function that takes the value 1 when the statement in brackets is true and 0 when is false; and β_k and σ_k are the parameters estimated. The pdf of this distribution looks like a triangle, and its main advantage is that it has a definite upper and lower limits resulting in shorter tails than the normal distribution. See for example panel B of Figure 2.2.

3. Uniform distribution:

$$\begin{aligned}\beta_{k,ir} &= \beta_k + \sigma_k(2 \times u_{k,ir} - 1)\\ u_{k,ir} &\sim U(0,1)\end{aligned}$$

where β_k and σ_k are estimated. In this case, the parameter for each individual is equally likely to take on any value in some interval (see panel C of Figure 2.2). Note also that the uniform distribution with bounds 0 and 1 is very suitable when there exists individual heterogeneity in a dummy variable. For this case, the restriction $\beta_k = \sigma_k = 1/2$ can be applied.

4. Log-Normal distribution:

$$\begin{aligned}\beta_{k,ir} &= \exp\left(\beta_k + \sigma_k w_{k,ir}\right) \\ w_{k,ir} &\sim N(0,1)\end{aligned}$$

where β_k and σ_k are estimated. Then, $\beta_{k,i} \sim \log N(\beta_k, \sigma_k^2)$. The support of the log-normal distribution is $(0, +\infty)$, therefore the coefficient is allowed to have individual heterogeneity in the positive domain only (see panel D of Figure 2.2). If the some coefficient is expected a priori to be negative for all the individuals, one can create the negative of the variable and then include this new variable in the estimation. This allows the coefficient to be negative without imposing a sign change in the estimation procedure (Train, 2009).

5. Truncated (at 0) normal distribution:

$$\begin{aligned}\beta_{k,ir} &= \begin{cases}\beta_k + \sigma_k w_{k,ir} & \text{if} \quad \beta_{k,ir} > 0 \\ 0 & \text{otherwise}\end{cases} \\ w_{k,ir} &\sim N(0,1)\end{aligned}$$

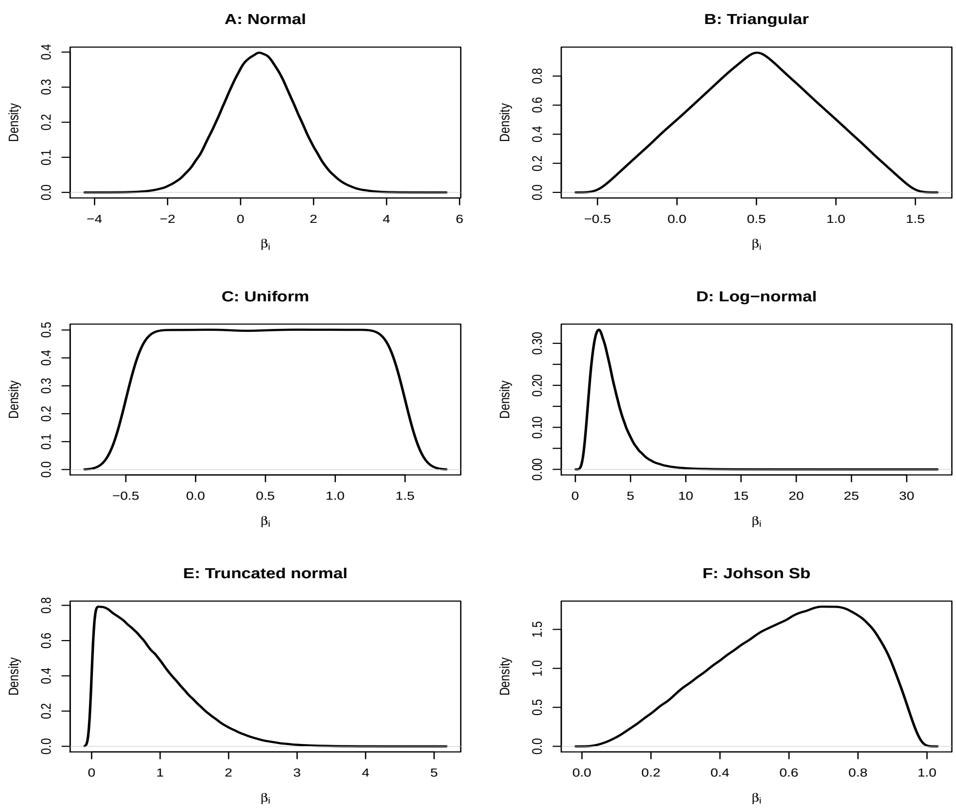
where β_k and σ_k are estimated. Then, $\beta_{k,i} \sim N(\beta_k, \sigma_k^2)$ with the share below zero massed at zero. This distribution is useful when the researcher has a priori belief that for some individuals the marginal latent effect of the variable is null. See panel F of Figure 2.2.

6. Johnson's S_b distribution:

$$\begin{aligned}\beta_{k,ir} &= \frac{\exp\left(\beta_k + \sigma_k w_{k,ir}\right)}{1 + \exp\left(\beta_k + \sigma_k w_{k,ir}\right)} \\ w_{k,ir} &\sim N(0,1)\end{aligned}$$

where β_k and σ_k are estimated. This distribution gives coefficients between 0 and 1, which is also very suitable for dummy variables. If the researcher needs the coefficient to be between 0 and d, then the variable can be multiplied by d before estimation. The main advantage of the Johnson S_b is that it can be shaped like log-normal distribution, but with thinner tails below the bound. See panel E of Figure 2.2.

Figure 2.2: Distributions for random coefficients.



Rchoice allows to the user to specify both types of random draws by the argument `haltons`: pseudo-random draws (`haltons = NULL`) and Halton draws (`haltons = NA`) as default. For the Halton draws, the default is to use the first K prime numbers starting with 3. Within each series, the first 100 draws are discarded, as the first draws tend to be highly correlated across different draws.

The user can also change the prime number and the number of elements dropped for each series. For example, if $K = 2$, and the user wants to use prime numbers 5 and 31 along with dropping the first 10 draws, he could specify `haltons = list("prime" = c(5,31), "drop" = c(10,10))`.

2.5 Applications using Rchoice

In this section I present some of the capabilities of **Rchoice**. Section 2.5.1 shows how to estimate Poisson, binary and ordered response models with fixed parameters. Examples of how to estimate models with random parameters are presented in Section 2.5.2. Section 2.5.3 explains how to estimate models with random parameters with panel or longitudinal data. In Section 2.5.4, models with both observed and unobserved heterogeneity are estimated. Finally, Section 2.5.5 shows how the conditional estimates for each individual can be plotted.

2.5.1 Standard models

In this section, I show the capabilities of **Rchoice** to estimate Poisson, binary and ordered regression models without random parameters. The main objective of this section is to show how **Rchoice** can interact with other packages in R.

To show how to estimate Poisson regression models using **Rchoice**, I will use data on scientific productivity (Long, 1990, 1997). We load the data using

```
R> data("Articles")
R> head(Articles,3)
```

```
  art fem mar kid5  phd ment
1   0   0   1    0 2.52    7
2   0   1   0    0 2.05    6
3   0   1   0    0 3.75    6
```

To see more information about the data, one can use:

```
R> help(Articles)
```

The work by Long (1990) suggests that gender, marital status, number of young children, prestige of the graduate program, and the number of articles written by a scientist's mentor could affect a scientist's level of publication. To see this, we estimate a Poisson regression model and use the `Rchoice` function specifying `family = poisson`:

```
R> poisson.fixed <- Rchoice(art ~ fem + mar + kid5 + phd + ment,
+                      data = Articles,
+                      family = poisson)
R> summary(poisson.fixed)

Model: poisson
Model estimated on: Mon Sep 07 15:16:54 2015

Call:
Rchoice(formula = art ~ fem + mar + kid5 + phd + ment, data = Articles,
    family = poisson, method = "nr")
```

```
The estimation took: 0h:0m:0s

Coefficients:
          Estimate Std. Error z-value Pr(>|z|)
constant  0.304617   0.102982   2.958   0.0031 **
fem      -0.224594   0.054614  -4.112 3.92e-05 ***
mar       0.155243   0.061375   2.529   0.0114 *
kid5     -0.184883   0.040127  -4.607 4.08e-06 ***
phd       0.012823   0.026397   0.486   0.6271
ment      0.025543   0.002006  12.733  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Optimization of log-likelihood by Newton-Raphson maximisation
Log Likelihood: -1651
Number of observations: 915
Number of iterations: 7
Exit of MLE: gradient close to zero
```

The output shows that the log-likelihood function is estimated using the Newton-Raphson algorithm in 7 iterations. In terms of interpretation, we can say that, being a female scientist decreases the expected number of articles by a factor of 0.8 ($= \exp(-.225)$), holding all other variables constant. Or equivalently, being a female scientist decreases the expected number of articles by 20% ($= 100\,[\exp(-.225) - 1]$), holding all other variables constant. Prestige of the PhD department is not important for productivity.

Another capability of **Rchoice** is its interaction with other packages in R. For example, we can compute the robust standard error by using the package **sandwich** (Zeileis, 2006):

```
R> library("sandwich")
R> library("lmtest")
R> coeftest(poisson.fixed, vcov = sandwich)

t test of coefficients:

           Estimate Std. Error t value  Pr(>|t|)
constant  0.3046168  0.1465197  2.0790 0.0378958 *
fem      -0.2245942  0.0716622 -3.1341 0.0017793 **
mar       0.1552434  0.0819292  1.8948 0.0584297 .
kid5     -0.1848827  0.0559633 -3.3036 0.0009917 ***
phd       0.0128226  0.0419642  0.3056 0.7600096
ment      0.0255427  0.0038178  6.6905 3.884e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To get the same robust standard errors as STATA (StataCorp, 2011), we need to make a small sample correction:

```
R> vcov.stata <- vcovHC(poisson.fixed, type = "HC0") *
+                nObs(poisson.fixed) / (nObs(poisson.fixed) - 1)
R> coeftest(poisson.fixed, vcov = vcov.stata)
```

```
t test of coefficients:

          Estimate Std. Error t value  Pr(>|t|)
constant  0.3046168  0.1465999  2.0779  0.038001 *
fem      -0.2245942  0.0717014 -3.1324  0.001790 **
mar       0.1552434  0.0819740  1.8938  0.058567 .
kid5     -0.1848827  0.0559939 -3.3018  0.000998 ***
phd       0.0128226  0.0419871  0.3054  0.760137
ment      0.0255427  0.0038198  6.6868 3.977e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

where the correction is $n/(n-1)$.

Rchoice also interacts with the `linearHypothesis` and `deltaMethod` functions from **car** (Fox et al., 2009) and the `lrtest` and `waldtest` functions from the **lmtest** package (Zeileis and Hothorn, 2002). For example, we can test the nonlinear hypothesis that the ratio between the `phd` and `ment` coefficient is zero[5], $H_0 : \texttt{phd/ment} = 0$, by:

```
R> library("car")
R> deltaMethod(poisson.fixed, "phd/ment")

          Estimate       SE
phd/ment 0.5020048 1.043031
```

[5]This ratio can be thought as a compensating variation between both variables. See for example Section 2.5.5 for further details of compensating variation with individual heterogeneity.

The main argument to estimate other models is `family` (see Table 2.1). This provides a convenient way to specify the details of the models used by `Rchoice`. For Probit models, the user should specify `family = binomial("probit")`, and for Logit `family = binomial("logit")`. In the following example, I use the `Workmroz` data base to estimate a binary Probit model, where the dependent variable `lfp` equals 1 if wife is in the paid labor force, and 0 otherwise.

```
R> data("Workmroz")
R> oprobit <- Rchoice(lfp ~ k5 + k618 + age + wc + hc + lwg + linc,
+                     data = Workmroz,
+                     family = binomial('probit'))
R> summary(oprobit)
```

```
Model: binomial
Model estimated on: Mon Sep 07 15:16:54 2015

Call:
Rchoice(formula = lfp ~ k5 + k618 + age + wc + hc + lwg + linc,
    data = Workmroz, family = binomial("probit"), method = "nr")



Frequencies of categories:
y
     0      1
0.4322 0.5678
The estimation took: 0h:0m:0s
```

```
Coefficients:
          Estimate Std. Error z-value Pr(>|z|)
constant  2.781983   0.441876   6.296 3.06e-10 ***
k5       -0.880688   0.113436  -7.764 8.22e-15 ***
k618     -0.038656   0.040454  -0.956 0.339304
age      -0.037701   0.007612  -4.953 7.31e-07 ***
wc        0.481150   0.135271   3.557 0.000375 ***
hc        0.077440   0.124733   0.621 0.534700
lwg       0.371645   0.087605   4.242 2.21e-05 ***
linc     -0.451494   0.100748  -4.481 7.42e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Optimization of log-likelihood by Newton-Raphson maximisation
Log Likelihood: -451.9
Number of observations: 752
Number of iterations: 4
Exit of MLE: gradient close to zero
```

Ordered Probit and Logit models are estimated in the same way. In this case we use the `Health` database and create the logarithm of household income. The dependent variable, `newhsat`, is a categorical variable that indicates the self reported health assessment of individuals recorded with values 0,1,...,4.[6]

```
R> data("Health")
R> Health$linc <- log(Health$hhinc)
```

[6]See `help(Health)`.

```
R> ologit <- Rchoice(newhsat ~ age + educ + married + hhkids + linc,
+                      data = head(Health, 2000),
+                      family = ordinal('logit'))
R> summary(ologit)

Model: ordinal
Model estimated on: Mon Sep 07 15:16:54 2015

Call:
Rchoice(formula = newhsat ~ age + educ + married + hhkids + linc,
    data = head(Health, 2000), family = ordinal("logit"), method = "bfgs")



Frequencies of categories:
y
     0      1      2      3      4
0.0600 0.2675 0.4545 0.1010 0.1170
The estimation took: 0h:0m:0s

Coefficients:
          Estimate Std. Error z-value Pr(>|z|)
kappa.1   2.094209   0.090487  23.144  < 2e-16 ***
kappa.2   4.198067   0.105952  39.623  < 2e-16 ***
kappa.3   4.961227   0.115706  42.878  < 2e-16 ***
constant  2.170194   0.778474   2.788  0.00531 **
age      -0.030637   0.004517  -6.782 1.19e-11 ***
```

```
educ      0.065848   0.018456   3.568  0.00036 ***
married  -0.336312   0.110278  -3.050  0.00229 **
hhkids    0.220423   0.099485   2.216  0.02672 *
linc      0.176496   0.098799   1.786  0.07403 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Optimization of log-likelihood by BFGS maximisation
Log Likelihood: -2663
Number of observations: 2000
Number of iterations: 67
Exit of MLE: successful convergence
```

2.5.2 Random parameters models with cross sectional data

The main advantage of **Rchoice** over other packages is that it allows estimating models with random parameters. In this section, I show how to estimate those kinds of models for count data using cross-sectional data. For binary and ordered models the syntax is the same provided that the `family` argument is correctly specified.

Continuing with the previous Poisson model, now I will assume that the effect of `kid5`, `phd` and `ment` are not fixed, but rather heterogeneous across the population. Specifically, I will assume that the coefficients for those variables are independently normally distributed, that is, I will not allow correlation among them. In particular,

$$
\begin{aligned}
\beta_{\texttt{kid5},ir} &= \beta_{\texttt{kid5}} + \sigma_{\texttt{kid5}}\omega_{\texttt{kid5},ir} \\
\beta_{\texttt{phd},ir} &= \beta_{\texttt{phd}} + \sigma_{\texttt{phd}}\omega_{\texttt{phd},ir} \\
\beta_{\texttt{ment},ir} &= \beta_{\texttt{ment}} + \sigma_{\texttt{ment}}\omega_{\texttt{ment},ir}
\end{aligned}
$$

where $\omega_{k,ir} \sim N(0,1)$. Then, the Poisson model with random parameter is estimated using the following syntax:

```
R> poisson.ran <- Rchoice(art ~ fem + mar + kid5 + phd + ment,
+                         data = Articles,
+                         family = poisson,
+                         ranp = c(kid5 = "n", phd = "n", ment = "n"))
```

It is important to discuss the arguments for the `Rchoice` function. First, the argument `ranp` indicates which variables are random in the `formula` and their distributions. In this case, I have specified that all of them are normal distributed using `"n"`. The shorthand for the remaining distributions that can be used are:[7]

- Log-Normal = `"ln"`,
- Truncated Normal = `"cn"`,
- Uniform = `"u"`,
- Triangular = `"t"`,
- Johnson's S_b = `"sb"`.

The number of draws are not specified. Therefore, `Rchoice` will set `R = 40` as default. The user can change this by changing the `R` argument. The type of

[7]These shorthands are the same as those used by **mlogit** and **gmnl** packages.

draws are Halton draws as default, but if users want pseudo-random draws they can specify `haltons = NULL` (see Section 2.4).

```
R> summary(poisson.ran)

Model: poisson
Model estimated on: Mon Sep 07 15:17:19 2015

Call:
Rchoice(formula = art ~ fem + mar + kid5 + phd + ment, data = Articles,
    family = poisson, ranp = c(kid5 = "n", phd = "n", ment = "n"),
    method = "bfgs", iterlim = 2000)

The estimation took: 0h:0m:25s

Coefficients:
           Estimate Std. Error z-value Pr(>|z|)
constant   0.225583   0.132500   1.703  0.08866 .
fem       -0.218498   0.070558  -3.097  0.00196 **
mar        0.156431   0.079121   1.977  0.04803 *
mean.kid5 -0.197775   0.063472  -3.116  0.00183 **
mean.phd  -0.029942   0.037217  -0.805  0.42109
mean.ment  0.031110   0.003814   8.158 4.44e-16 ***
sd.kid5    0.285310   0.089104   3.202  0.00136 **
sd.phd     0.165405   0.016585   9.973  < 2e-16 ***
sd.ment    0.015876   0.003535   4.491 7.11e-06 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Optimization of log-likelihood by BFGS maximisation
Log Likelihood: -1574
Number of observations: 915
Number of iterations: 81
Exit of MLE: successful convergence
Simulation based on 40 Halton draws
```

It is important to check the exit of the estimation. In our example, the output informs us that the convergence was achieved successfully. The results also show that the standard deviations of the coefficients are highly significant, indicating that parameters do indeed vary in the population. Since the parameters are normally distributed, we can also say that:

```
R> pnorm(coef(poisson.ran)["mean.kid5"]/coef(poisson.ran)["sd.kid5"])

mean.kid5
0.2440946
```

24% of the individuals have a positive coefficient for `kid5`. In other words, for about 76% of PhD students, having children less than 6 years old reduces their productivity. Note also that the mean coefficient for `phd` is 0 (not significant). This is due to the fact that the unobserved heterogeneity among scientists in the sample cancel out positive and negative effects. These observations are not possible with a Poisson regression with fixed parameters.

Suppose that now we want to test whether $H_0 = \sigma_{\texttt{kid5}} = \sigma_{\texttt{phd}} = \sigma_{\texttt{ment}} = 0$. This can be done by using the function `waldtest` or `lrtest` from package **lmtest**:

```
R> waldtest(poisson.fixed, poisson.ran)

Wald test

Model 1: art ~ fem + mar + kid5 + phd + ment
Model 2: art ~ fem + mar + kid5 + phd + ment
  Res.Df Df  Chisq Pr(>Chisq)
1    909
2    906  3 280.14  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R> lrtest(poisson.fixed, poisson.ran)

Likelihood ratio test

Model 1: art ~ fem + mar + kid5 + phd + ment
Model 2: art ~ fem + mar + kid5 + phd + ment
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   6 -1651.1
2   9 -1574.2  3 153.78  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both tests reject the null hypothesis. We can also specify different distribution of the parameters by using the S3 method `update`:

```
R> poisson.ran2 <- update(poisson.ran,
+                          ranp = c(kid5 = "u", phd = "t" , ment = "cn"))
```

Both models `poisson.ran` and `poisson.ran2` can be compared using `mtable` from **memisc** (Elff, 2014) package:

```
R> library("memisc")
R> Ttable <- mtable("model 1"= poisson.ran, "model 2" = poisson.ran2,
+          summary.stats = c("N", "Log-likelihood", "BIC", "AIC"))

R> Ttable

Calls:
model 1: Rchoice(formula = art ~ fem + mar + kid5 + phd + ment,
    data = Articles,
    family = poisson, ranp = c(kid5 = "n", phd = "n", ment = "n"),
    method = "bfgs", iterlim = 2000)
model 2: Rchoice(formula = art ~ fem + mar + kid5 + phd + ment,
    data = Articles,
    family = poisson, ranp = c(kid5 = "u", phd = "t", ment = "cn"),
    method = "bfgs", iterlim = 2000)

===================================
                 model 1   model 2
-----------------------------------
```

```
constant          0.226     0.254
                 (0.132)   (0.132)
fem              -0.218**  -0.216**
                 (0.071)   (0.069)
mar               0.156*    0.149
                 (0.079)   (0.076)
mean.kid5        -0.198**  -0.221***
                 (0.063)   (0.064)
mean.phd         -0.030    -0.104*
                 (0.037)   (0.041)
mean.ment         0.031***  0.028***
                 (0.004)   (0.003)
sd.kid5           0.285**   0.508***
                 (0.089)   (0.123)
sd.phd            0.165***  0.226***
                 (0.017)   (0.025)
sd.ment           0.016***  0.022***
                 (0.004)   (0.003)
-----------------------------------
N                  915       915
Log-likelihood   -1574.166 -1575.816
BIC               3209.702  3213.003
AIC               3166.332  3169.632
===================================
```

The previous model specifies the coefficients to be independently distributed while one would expect correlation. For example, the effect of the prestige of PhD department could be positively correlated with the number of publications by a mentor. Now, I estimate the model `poisson.ran`, but assuming that the random parameters are correlated: $\boldsymbol{\beta}_i \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ for a general matrix $\boldsymbol{\Sigma}$. The main argument for this model is `correlation = TRUE`:

```
R> poissonc.ran <- Rchoice(art ~ fem + mar + kid5 + phd + ment,
+                            data = Articles,
+                            ranp = c(kid5 = "n", phd = "n",
ment = "n"),
+                            family = poisson,
+                            correlation =  TRUE)
R> summary(poissonc.ran)

Model: poisson
Model estimated on: Mon Sep 07 15:18:49 2015

Call:
Rchoice(formula = art ~ fem + mar + kid5 + phd + ment,
     data = Articles,
    family = poisson,
    ranp = c(kid5 = "n", phd = "n", ment = "n"),
    correlation = TRUE, method = "bfgs", iterlim = 2000)

The estimation took: 0h:0m:57s

Coefficients:
              Estimate Std. Error z-value Pr(>|z|)
constant      0.235301   0.131432   1.790 0.073409 .
fem          -0.228057   0.070992  -3.212 0.001316 **
```

```
mar           0.150374   0.079625   1.889 0.058954 .
mean.kid5    -0.229971   0.063024  -3.649 0.000263 ***
mean.phd     -0.032431   0.037128  -0.874 0.382384
mean.ment     0.033804   0.003751   9.012  < 2e-16 ***
sd.kid5.kid5  0.279620   0.091789   3.046 0.002317 **
sd.kid5.phd   0.084343   0.055691   1.514 0.129908
sd.kid5.ment -0.025400   0.005943  -4.274 1.92e-05 ***
sd.phd.phd   -0.143787   0.028258  -5.088 3.61e-07 ***
sd.phd.ment  -0.002123   0.007752  -0.274 0.784153
sd.ment.ment  0.011351   0.007372   1.540 0.123616
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Optimization of log-likelihood by BFGS maximisation
Log Likelihood: -1571
Number of observations: 915
Number of iterations: 211
Exit of MLE: successful convergence
Simulation based on 40 Halton draws
```

The output prints the mean of the random parameters along with the lower-triangular Cholesky factor $\mathbf{L}$. We can extract the variance-covariance matrix, $\mathbf{\Sigma}$, and the correlation matrix of the random parameters using S3 method `vcov` in the following way:

```
R> vcov(poissonc.ran, what = "ranp", type = "cov")
```

```
             kid5          phd          ment
kid5  0.07818737  0.023583914 -0.0071022904
phd   0.02358391  0.027788312 -0.0018369769
ment -0.00710229 -0.001836977  0.0007785002
```

```
R> vcov(poissonc.ran, what = "ranp", type = "cor")
```

```
            kid5        phd      ment
kid5  1.0000000  0.5059604 -0.910334
phd   0.5059604  1.0000000 -0.394951
ment -0.9103340 -0.3949510  1.000000
```

The argument `what` indicates which covariance matrix has to be extracted. The default is `coefficient`, and the `vcov` behaves as usual. If `what = "ranp"` the covariance matrix of the random parameters is returned as default. The argument `type` indicates what matrix of the random parameters should be returned. Among other things, the output shows that `ment` is negatively related with `phd` and `kid5`. Specifically, we can see that the correlation between `phd` and `ment` is around -0.4. We can also test whether the variances of the random parameters are significant using Delta Method. To do so, we can use the `se = TRUE` argument in the `vcov` function, which is a wrapper of `deltamethod` function from **msm** (Jackson, 2011) package:

```
R> vcov(poissonc.ran, what = "ranp", type = "cov", se = TRUE)
```

```
Elements of the variance-covariance matrix
```

```
             Estimate  Std. Error z-value Pr(>|z|)
v.kid5.kid5  0.07818737  0.05133227  1.5232 0.127718
v.kid5.phd   0.02358391  0.01152073  2.0471 0.040650 *
v.kid5.ment -0.00710229  0.00245237 -2.8961 0.003778 **
v.phd.phd    0.02778831  0.00896418  3.0999 0.001936 **
v.phd.ment  -0.00183698  0.00186657 -0.9841 0.325044
v.ment.ment  0.00077850  0.00032116  2.4240 0.015349 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the covariance between `ment` and `phd` even though is negative is not significant. To get the standard errors of the standard deviations for the random parameters, we might use:

```
R> vcov(poissonc.ran, what = "ranp", type = "sd", se = TRUE)

Standard deviations of the random parameters

      Estimate Std. Error z-value  Pr(>|z|)
kid5 0.2796201  0.0917893  3.0463  0.002317 **
phd  0.1666983  0.0268875  6.1999 5.652e-10 ***
ment 0.0279016  0.0057552  4.8481 1.247e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.5.3 Random parameters models with panel data

The current version of this package also handles panel data by estimating random effect (RE) models. Suppose that the i observation $\mathbf{y}_i$ has unconditional joint density $f(\mathbf{y}_i^*|\mathbf{X}_i, \alpha_i, \boldsymbol{\beta})$ and the random effect has density $\alpha_i \sim g(\alpha_i|\boldsymbol{\theta})$, where $g(\alpha_i|\boldsymbol{\theta})$ does not depend on observables. Thus, the unconditional joint density for the ith observation is

$$f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\theta}) = \int_{\alpha_i} \left[\prod_{t=1}^{T_i} f(y_{it}^*|\mathbf{x}_{it}, \alpha_i, \boldsymbol{\beta})\right] g(\alpha_i|\boldsymbol{\theta}) d\alpha_i.$$

This model can be estimated using the package **pglm** (Croissant, 2013), which uses **Gauss-Hermite quadrature** to approximate the integration in the probability. However, note that this model can be seen as a random parameter model where the constant is random. Therefore, users might estimate a simple RE model with **Rchoice** by typing `ranp = (constant = 'n')`.

In the following example I will estimate a binary Probit model with RE and random parameters using `Unions` database from the **pglm** package.

```
R> data("Unions", package = "pglm")
R> Unions$lwage <- log(Unions$wage)
```

The model is estimated using the following syntax:

```
R> union.ran <- Rchoice(union ~ age + exper + rural + lwage,
+                       data = head(Unions, 2000),
+                       family = binomial('probit'),
```

```
+                              ranp = c(constant = "n", lwage = "t"),
+                              R = 10,
+                              panel = TRUE,
+                              index = "id",
+                              print.init = TRUE)
```

In this case, I assumed that `lwage` is distributed as triangular, while the constant is assumed to be normally distributed. This is the same as assuming that $\alpha_i \sim N(0, \sigma^2_\alpha)$.

There are two main arguments for the panel estimation. The argument `panel = TRUE` indicates that the data are a panel. This implies that users should indicate the variable that corresponds to the 'id' of the individuals in the `index` argument. In the this example, the 'id' is given by the variable "`id`".

Finally, the argument `print.init = TRUE` indicates that the initial values used by `Rchoice` will be displayed. This argument is useful if one wants to see the magnitude of the initial values for the parameters.

```
R> summary(union.ran)

Model: binomial
Model estimated on: Mon Sep 07 15:19:29 2015

Call:
Rchoice(formula = union ~ age + exper + rural + lwage,
    data = head(Unions, 2000),
    family = binomial("probit"), ranp = c(constant = "n",
```

```
    lwage = "t"), R = 10, panel = TRUE, index = "id",
    print.init = TRUE,
    method = "bfgs", iterlim = 2000)


Frequencies of categories:
y
     0      1
0.7605 0.2395
The estimation took: 0h:0m:40s

Coefficients:
              Estimate Std. Error z-value Pr(>|z|)
age            0.05407    0.32802   0.165    0.869
exper         -0.09174    0.02098  -4.372 1.23e-05 ***
ruralyes       0.22735    0.17537   1.296    0.195
mean.constant -1.30073    0.24497  -5.310 1.10e-07 ***
mean.lwage     0.01089    0.15419   0.071    0.944
sd.constant    1.19918    0.16110   7.444 9.79e-14 ***
sd.lwage       1.04202    0.10745   9.698  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Optimization of log-likelihood by BFGS maximisation
Log Likelihood: -738.6
Number of observations: 2000
```

```
Number of iterations: 77
Exit of MLE: successful convergence
Simulation based on 10 Halton draws
```

The results indicate that σ_α =1.2 and is significant. Furthermore, the finding of a significant standard deviation yet insignificant mean for `lwage` attests to the existence of substantial heterogeneity; positive and negative coefficients in the sample compensate for each other, such that the coefficient of the mean is not significant.

As in the previous cases, an ordered Probit Model with RE and random parameters can be estimated in the same way, but changing the distribution with the `family` argument:

```
R> oprobit.ran <- Rchoice(newhsat ~ age + educ + married +
                          hhkids + linc,
+                         data = head(Health, 2000),
+                         family = ordinal('probit'),
+                         ranp = c(constant = "n", hhkids = "n",
+                                  linc = "n"),
+                         panel = TRUE,
+                         index = "id",
+                         R = 100,
+                         print.init = TRUE)
R> summary(oprobit.ran)
```

2.5.4 Random parameter model with observed heterogeneity

In this section I illustrate how to estimate a Poisson random parameter model with observed heterogeneity. In the following example, I assume that there exists not only unobserved heterogeneity in the coefficients for `phd` and `ment`, but also observed heterogeneity in the mean of the coefficients. Specifically, I assume that the coefficients vary across individuals according to:

$$
\begin{aligned}
\beta_{\texttt{kid5},ir} &= \beta_{\texttt{kid5}} + \sigma_{\texttt{kid5}}\omega_{\texttt{kid5},ir} \\
\beta_{\texttt{phd},ir} &= \beta_{\texttt{phd}} + \pi_{\texttt{phd},\texttt{fem}}\texttt{fem} + \sigma_{\texttt{phd}}\omega_{\texttt{phd},ir} \\
\beta_{\texttt{ment},ir} &= \beta_{\texttt{ment}} + \pi_{\texttt{ment},\texttt{fem}}\texttt{fem} + \pi_{\texttt{ment},\texttt{phd}}\texttt{phd} + \sigma_{\texttt{ment}}\omega_{\texttt{ment},ir}.
\end{aligned}
$$

The formulation above implies that, for example, the `ment`'s coefficient (or the marginal effect on latent productivity) varies by gender and `phd`.

```
R> poissonH.ran <- Rchoice(art ~ fem + mar + kid5 + phd
                           + ment|fem + phd,
+                          data = Articles,
+                          ranp = c(kid5 = "n", phd = "n",
                                  ment = "n"),
+                          mvar = list(phd = c("fem"),
+                                      ment = c("fem", "phd")),
+                          family = poisson,
+                          R = 10)
```

Rchoice manages the variables in the hierarchical model by the `formula` object. Note that the second part of the `formula` is reserved for all the variables that enter in the mean of the random parameters. The argument `mvar` is a list that indicates how all this variables enter in each random parameter. For example `phd = c("fem")` indicates that the mean of `phd` coefficient varies according to `fem`.

```
R> summary(poissonH.ran)

Model: poisson
Model estimated on: Mon Sep 07 15:20:26 2015

Call:
Rchoice(formula = art ~ fem + mar + kid5 + phd + ment | fem +
    phd, data = Articles, family = poisson, ranp = c(kid5 = "n",
    phd = "n", ment = "n"), R = 10, mvar = list(phd = c("fem"),
    ment = c("fem", "phd")), method = "bfgs", iterlim = 2000)

The estimation took: 0h:0m:57s

Coefficients:
           Estimate Std. Error z-value Pr(>|z|)
constant   0.222646   0.181920   1.224  0.22100
fem       -0.572068   0.217337  -2.632  0.00848 **
mar        0.176966   0.075165   2.354  0.01855 *
mean.kid5 -0.259958   0.061981  -4.194 2.74e-05 ***
mean.phd  -0.019339   0.056808  -0.340  0.73354
mean.ment  0.048094   0.009960   4.829 1.37e-06 ***
```

```
phd.fem    0.133917   0.068995   1.941  0.05226 .
ment.fem  -0.006846   0.007134  -0.960  0.33721
ment.phd  -0.004857   0.002902  -1.674  0.09415 .
sd.kid5    0.431139   0.085080   5.067 4.03e-07 ***
sd.phd     0.133125   0.015333   8.682  < 2e-16 ***
sd.ment    0.014867   0.003050   4.874 1.09e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Optimization of log-likelihood by BFGS maximisation
Log Likelihood: -1577
Number of observations: 915
Number of iterations: 99
Exit of MLE: successful convergence
Simulation based on 10 Halton draws
```

The estimated parameters indicate that gender matters only for `phd`'s mean coefficient. Moreover, the results indicate that the prestige of the PhD reduce the effect of `ment`. We can test whether the interaction variables are jointly significant by using `lrtest` function:

```
R> lrtest(poisson.ran, poissonH.ran)

Likelihood ratio test
Model 1: art ~ fem + mar + kid5 + phd + ment
Model 2: art ~ fem + mar + kid5 + phd + ment | fem + phd
  #Df  LogLik Df  Chisq Pr(>Chisq)
```

```
1   9 -1574.2
2  12 -1576.9  3 5.4385     0.1424
```

2.5.5 Plotting conditional means

It is important to note that the estimates of the model parameters provide the unconditional estimates of the parameter vector, but we can derive an individual-specific conditional estimator (see Train, 2009; Greene, 2012). The estimator of the conditional mean of the distribution of the random parameters, conditioned on the person-specific data, is:[8]

$$\widehat{\widehat{\boldsymbol{\beta}}}_i = \widehat{\mathbb{E}}(\boldsymbol{\beta}_i|\text{data}_i) = \sum_{r=1}^{R}\left(\frac{\widehat{P}(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_i)}{\sum_{r=1}^{R}\widehat{P}(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_i)}\right)\widehat{\boldsymbol{\beta}}_{ir},$$

where:

$$\widehat{\boldsymbol{\beta}}_{ir} = \widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Pi}}\mathbf{s}_i + \widehat{\mathbf{L}}\boldsymbol{\omega}_{ir}.$$

Note that these are not actual estimates of $\boldsymbol{\beta}_i$, but are estimates of the conditional mean of the distribution of the random parameters (Greene et al., 2014). We can also estimate the standard deviation of this distribution by estimating:

$$\widehat{\mathbb{E}}(\boldsymbol{\beta}_i^2|\text{data}_i) = \sum_{r=1}^{R}\left(\frac{\widehat{P}(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_i)}{\sum_{r=1}^{R}\widehat{P}(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_i)}\right)\widehat{\boldsymbol{\beta}}_{ir}^2,$$

and then computing the square root of the estimated variance:

[8] Note that this simulator is also known as the posterior distribution of the individual parameters.

$$\sqrt{\widehat{\mathbb{E}}(\boldsymbol{\beta}_i^2|\text{data}_i) - \widehat{\mathbb{E}}(\boldsymbol{\beta}_i|\text{data}_i)^2}.$$

With the estimates of the conditional mean and conditional variance, we can then compute the limits of an interval that resembles a confidence interval as the mean plus and minus two estimated standard deviation. This will construct an interval that contains at least 95 percent of the conditional distribution of $\boldsymbol{\beta}_i$ (Greene, 2012).

Rchoice allows plotting the histogram and kernel density of conditional means of random parameters using the function `plot`. For instance, the kernel of the conditional mean of $\beta_{\texttt{lwage},i}$ for `union.ran` model can be obtained by typing:

```
R> plot(union.ran, par = "lwage", type = "density")
```

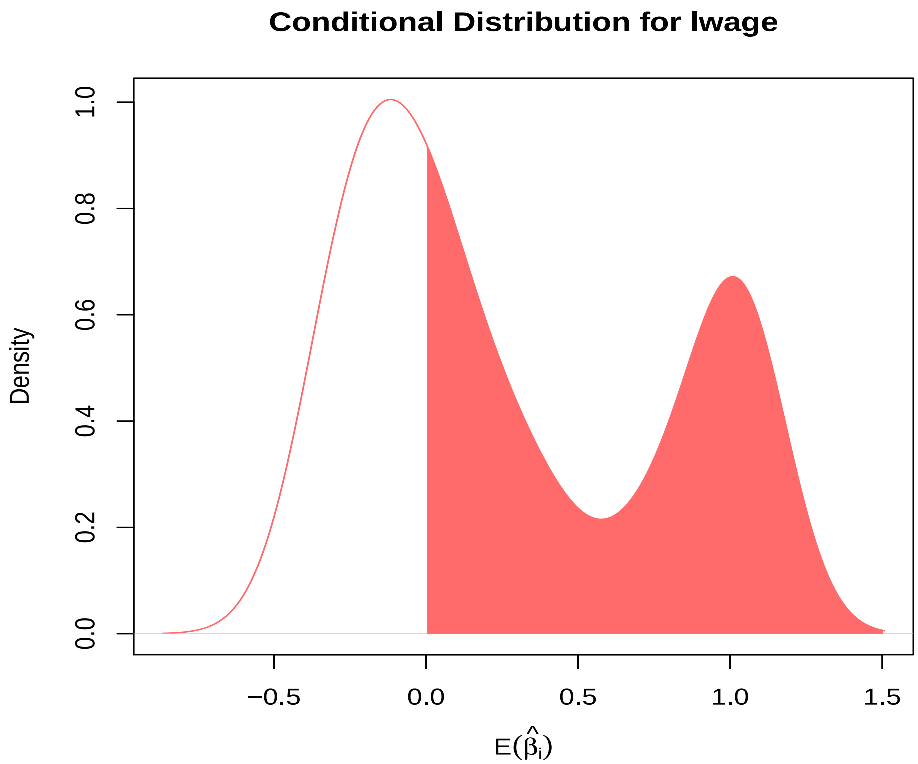
The graph produced using this command is visualized in Figure 2.3. We may also plot the individual confidence interval for the conditional means for the first 20 individuals using:

```
R> plot(union.ran, par = "lwage", ind =  TRUE, id = 1:20,
        col = "blue")
```

Figure 2.4 displays the individual conditional means and their respective confidence interval.

The method `plot` for objects of class `Rchoice` is a wrapper of `effect.Rchoice` function. This function retrieves the individual's conditional mean of both the parameters and the compensating variations. For example, one can get the individual's conditional mean and standard errors plotted in Figure 2.3 typing:

Figure 2.3: Kernel density of the individual's conditional mean.

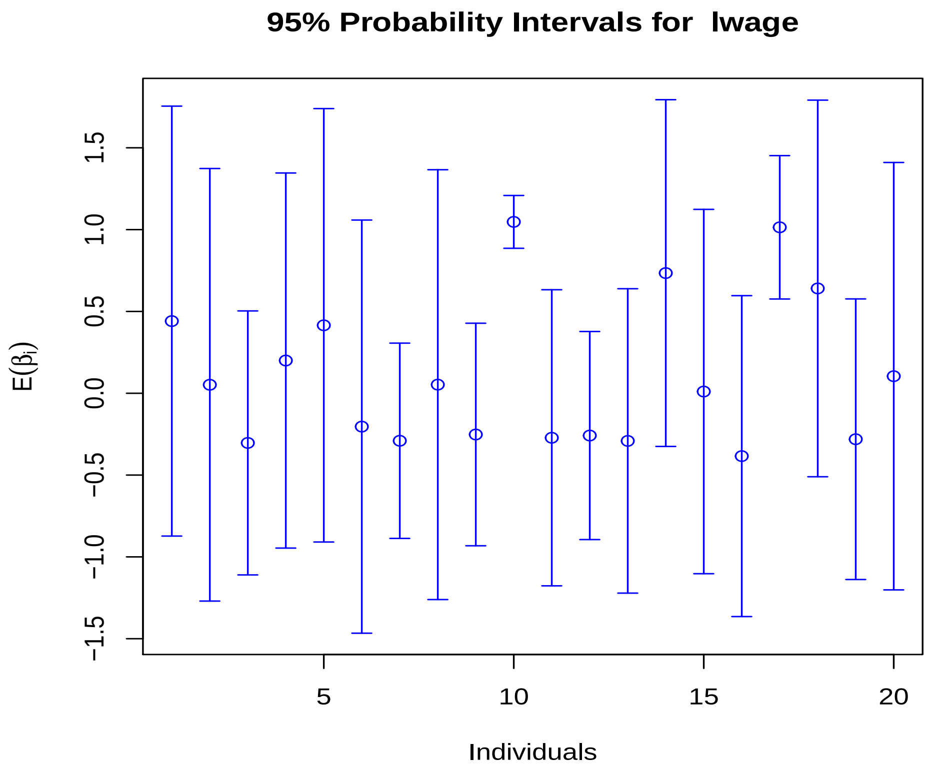



```
R> bi.wage <- effect.Rchoice(union.ran, par = "lwage",
              effect = "ce")
```

The argument `effect` is a string indicating what conditional mean should be computed. In this example, we are requiring the conditional expectation of the individual coefficients `effect = "ce"`. `effect.Rchoice` returns two list. The first one with the estimated conditional means for all the individuals, and the second one with the estimated standard errors of the conditional means.

```
R> summary(bi.wage$mean)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.4116 -0.1460  0.0718  0.2824  0.8643  1.0530
```

Figure 2.4: Individual confidence interval for the conditional means.




```
R> summary(bi.wage$sd.est)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0000017 0.3440000 0.5372000 0.4526000 0.6290000 0.8440000
```

One might also get the individual's conditional mean of the **compensating variations** using both `plot` and `effect.Rchoice`. Compensating variation is the variation in two regressors such that the latent variable does not change. Let x_{il} denote the lth elment in $\mathbf{x}_i$ and β_l the corresponding parameter, and let m index the mth elements in both vectors $\mathbf{x}_i$ and $\boldsymbol{\beta}$, respectively. Now consider a change in x_{il} and x_{im} at the same time, such that $y_i^* = 0$. This requires

$$0 = \beta_l \Delta x_{il} + \beta_{im} \Delta x_{im} \implies \frac{\Delta x_{il}}{\Delta x_{im}} = -\frac{\beta_{im}}{\beta_l}$$

where β_{im} is a random coefficient. This ratio (without the minus sign) is computed or plotted if the argument `effect = "cv"` in any of the two functions. The argument `par` is the variable whose coefficient goes in the numerator (β_{im}), and the argument `wrt` is a string indicated which coefficient goes in the denominator (β_l). Note that since β_{im} is random, the ratio of the coefficient is random and its distribution follows from the joint distribution of the coefficients.

2.6 Computational issues

The estimated parameters in any model estimated using SML depend on at least four factors. The first of them is the random number seed. If the random draws used in the estimation are pseudo-random draws, instead of Halton draws, then the parameter might change if the seed is changed. As default, **Rchoice** sets `seed = 10`. Second, the number of draws is very important for the asymptotic properties of the SML (see Section 3.3.2). In the examples provided in this document, I used just a few draws for time restrictions. However, in real applied work, researchers must use a greater number of draws, especially if pseudo-random draws are used in the estimation. A good starting point is 500 draws (see for example Bhat, 2001; Sándor and Train, 2004).

Starting values are also crucial to achieve convergence, specially for the correlated random parameter model with deterministic variation in the mean. Note that version 0.1 of **Rchoice** sets the initial values at 0, while versions 0.2 and 0.3 set them at 0.1 as default. Users might change this with the argument `init.ran` or setting a new vector of starting values with the argument `start`.

Finally, the choice of optimization method is another important factor that influences model convergence. As explained before, **Rchoice** uses the function `maxLik` in order to maximize the log-likelihood function, which permits one to estimate models by the Newton-Raphson (NR), BGFS and Berndt-Hall-Hall-Hausman (BHHH) procedures. I have found that BFGS often works best in the sense it is more likely to converge than the alternative algorithm. That is the main reason why this algorithm is set as default. If this method does not converge, users might re-estimate the model using another algorithm. For example, the user might type `method = "nr"` for the NR method or `method = "bhhh"` for the BHHH method in the `Rchoice` function. `Rchoice` uses the numerical Hessian if `method = "nr"` and the model is estimated with random parameters, thus it can be very slow compared to the other methods. It is worth mentioning that BHHH is generally faster than the other procedures, but it can failure if the variables have very different scale. The larger the ratio between the largest standard deviation and the smallest standard deviation, the more problems the user will have with the estimation procedure. Given this fact, users should check the variables and re-scale or recode them if necessary. Furthermore, it is also convenient to use the argument `print.level = 2`, for example, to trace the optimization procedure in real time. For more information about arguments for optimization type `help(maxLik)`.

2.7 Conclusions

The **Rchoice** package contains most of the newly developed models in binary, count and ordered models with random parameters. The current version of **Rchoice** handles cross-sectional and panel data with observed and unobserved individual heterogeneity. Allowing parameter values to vary across the population according to some pre-specified distribution overcomes the problem of having a fixed-representative coefficient for all individuals. The distributions supported by **Rchoice** are normal, log-normal, uniform, truncated normal, triangular distribution and Johnson's S_b. It also allows the user to choose between Halton draws and pseudo-random numbers and correlated parameters.

The **Rchoice** package intends to make available those estimation methods to the general public and practitioners in a friendly and flexible way. In future versions, I expect to add functions that allow estimating latent class models. I also hope to include functions to compute marginal effects.

CHAPTER 3

DO MONETARY SUBJECTIVE WELL-BEING EVALUATIONS VARY ACROSS SPACE? COMPARING CONTINUOUS AND DISCRETE SPATIAL HETEROGENEITY

3.1 Introduction

Subjective well-being (SWB) measures have been widely used in different fields, with particular interest in economics (Frey and Stutzer, 2002; Bruni and Porta, 2005; Bruni, 2007; Frey, 2008) and sociology (Eid and Larsen, 2008; Deiner et al., 1999). These measures allow researchers to obtain the impact of certain variables on a direct measure of welfare by estimating 'SWB or happiness equations' (Powdthavee, 2010b). The idea is to use self-rated happiness or life satisfaction measures, and to model individuals' SWB as a function of socio-economic characteristics. The resulting coefficients are the estimated average weights of each attribute on happiness.[1]

There is also a growing applied research line that uses self-reported well-being variables for evaluating compensating variation (CV). This method allows one to value in monetary terms a change in a certain variable (or condition) on individual's welfare, especially for attributes that do not have a market price (Clark and Oswald, 2002). This particular feature makes the CV method an attractive tool for policy makers that would like to know how individuals value intangible sources. For instance, CV method using SWB measures has been used for evaluating the monetized value of health impairments, in particular cardiovascular disease (Groot

[1]For a more complete review of how these measures have been used in the field in economic see Di Tella and MacCulloch (2006).

and Maassen van den Brink, 2006) and chronic diseases (Ferrer-i Carbonell and van Praag, 2002). It has been also been used to compute the value of social relationships (Powdthavee, 2008), crime (Moore and Shepherd, 2006), and for computing monetary values of externalities such as noise damage caused by aircraft nose nuisance (Van Praag and Baarsma, 2005) and other environmental conditions (Welsch and Kühling, 2009; Levinson, 2012).

In essence, CV computes the trade-off between income and the variable of interest by taking the ratio of the estimated parameters. The resulting ratio is called the CV, which computes the increase or decrease in income necessary to compensate individuals for any given change in some attribute or situation. This simplicity makes this method very appealing compared with other methods such as the hedonic price method or the contingent valuation method, in which people are asked directly their willingness to pay for some attributes. As a matter of fact, one of the main advantages of the SWB approach is that it permits the monetization of virtually any attribute or characteristic regardless of whether affected individuals are consciously aware of the effect or not (Welsch and Kühling, 2009). For a detailed review of the differences between these methods see for example Levinson (2012).

However, one issue that has been neglected in previous studies is the spatial heterogeneity that might be present in how individuals value some attributes or situations (and hence in CVs). Previous studies compute the CV using 'global models' (as opposed to local models) that assume that the estimated coefficients (and hence the CVs) are unique for all spatial units in the sample.[2] This approach

[2]Specifically, any model that estimates average relationships between the dependent and independent variables are considered global models. These models have the implicit assumption that variation is the same everywhere. Whereas local methods take into account the potential spatial variations by estimating a specific coefficient for each spatial unit. See for example Lloyd (2010) for a review of different methods used to capture spatial heterogeneity.

might be suitable if the goal is to obtain and test average CVs for the whole sample or establish stylized facts, regardless of the spatial location[3] of the individuals. However, one main drawback is that these methods may hide significant spatial variation and important local differences in the relationship between well-being and individual's characteristics resulting in misleading and inadequate policy inferences for spatial units that follow a completely different process than the average pattern (Ali et al., 2007).

This problem is particularly important for policy-making and cost-benefit analysis in problems that involve valuation of non-market attributes. For example, the government may be interested in how individuals value in monetary terms the negative externalities of noise pollution in their neighborhood, so as to create policies that compensate for the loss of welfare. Now imagine that a certain study indicates that individuals, regardless of their location, should be compensated say in average by US$50 in order to let them reach his previous welfare without the noise pollution. However, inhabitants in region A may be more sensitive to noise to those in region B due to observed and/or unobserved factors. Suppose that given this spatial heterogeneity in the sensitivities, individuals living in region A should be compensated by US$80, and those in region B by US$50. As a result, a spatially blind policy based on a unique estimate of CV for all regions may ignore this fact creating a welfare-gap of US$30 not covered for inhabitants in region A, and thus producing undesired geographically uneven impacts.

Naturally, a question that arises is: why compensating variations might vary across geographical space? According to Fotheringham et al. (2003) there are at least three reasons. The first one, and less interesting, is sampling variation, which relates to a statistical artifact and not to any underlying spatial process. A second

[3] Throughout this study I will use location unit, region, or geographical area interchangeably.

potential cause is misspecification or important variables that follow a spatial non-stationary process but are omitted from the model. The third possibility is that people's preferences for some attributes are intrinsically different across space. That is, individuals may differentiate strongly in terms of valuation of changes in characteristics among some locations or areas producing different reactions to the same stimuli over space.

Under this context, the aim of this chapter is to address the limitation of previous studies by taking into consideration the spatial heterogeneity that might be present in individuals' evaluations, where spatial heterogeneity is understood as different realizations of the parameters for each spatial unit. To accomplish this, I describe and compare two econometric ways of modeling unobserved spatial heterogeneity. The idea in both approaches is to allow the coefficients to vary across spatial units by assuming some distribution specified a priori. By allowing spatially varying coefficients I will be able to obtain location-specific CVs. The main difference between both approaches is whether the chosen distribution is continuous or discrete. The hypothesis of spatial heterogeneity in CVs is tested using a Chilean database with communes as spatial units.[4]

I present two main results. First, both approaches reveal that there exist significant amounts of spatial variation on individuals' valuations for some attributes. This is a very important result for the literature on happiness and economics: neglecting spatial heterogeneity an ignoring local differences on CVs might lead to inadequate policy design for some group of regions. In effect, if different regions have different valuations of a certain observable attribute, then one should account for this heterogeneity when analyzing the potential outcomes of a policy. The second important result is that the discrete approach provides better statistical fit and

[4]Communes are the smallest geographical areas in Chile for which individual data is available.

easier interpretation of the varying compensating variations than the continuous approach. This might be explained by the complexity of the estimation procedure when assuming continuous unobserved spatial heterogeneity.

The remainder of this chapter is organized as follows. Section 3.2 discusses a simple model used to understand spatial heterogeneity in CVs, along with its underlying assumptions. Section 3.3 explains the continuous and discrete strategies for modeling spatial heterogeneity. Section 3.4 presents the data and results. Finally, Section 3.5 discusses the main findings and concludes.

3.2 Model, assumptions and spatial heterogeneity

3.2.1 Model and assumptions

Similarly to Ferrer-i Carbonell and van Praag (2002) and Welsch and Kühling (2009), I assume that the true latent well-being of individuals is determined by income Y, specific health conditions H, socio-economic X characteristics, and neighborhood attributes Z

$$W_i^* = W(Y_i, H_i, X_i, Z_i). \tag{3.1}$$

This latent function can be thought of as an operationalization of an indirect utility function, which gives the individuals' maximal attainable utility when faced with those characteristics (see for example Van Praag and Baarsma, 2005; Bockstael and Freeman, 2005). Adopting this notation, one could ask for example: *what is the change in income needed to compensate individuals for the marginal change* *of neighborhood characteristics?* Mathematically, this is achieved by totally differentiating Equation (3.1) in equilibrium, setting $dW^* = 0$, and holding all the other variables constant such that:

$$CV = \left.\frac{\partial Y}{\partial Z}\right|_{dW^*=0} = -\frac{\frac{\partial W}{\partial Z}}{\frac{\partial W}{\partial Y}}. \tag{3.2}$$

The ratio in Equation (3.2) measures the marginal change in Y needed to bring the individual to his original level of well-being given a marginal change in Z, *ceteris paribus*. Due to the negative sign, CV will be negative if Z measures good quality of the neighborhood and positive if it measures a bad characteristic, holding other things equal.[5]

A similar analysis can be done if the Z is an indicator variable. Consider an individual who experiences a loss in his welfare due to a negative externality in his neighborhood. Someone with an externality ($Z = 1$) enjoys the same welfare as someone without this externality ($Z = 0$) if

$$W(Y, H, X, Z = 0) = W(Y + \Delta Y, H, X, Z = 1). \tag{3.3}$$

In order to operationalize Equation (3.1), I proxied W_i^* by self-reported subjective well-being using the answer to the following question "*How satisfied are you with life?*" The answer ranges from 1 ("completely unsatisfied") to 10 ("completely satisfied"), and then I use a dichotomy version of this categorical answer. According to Ferrer-i Carbonell and Frijters (2004), this approach has three main theoretical assumptions.

[5]The main assumption is that greater income is associated with greater happiness. This assumption is extended in this work and empirically corroborated.

- *Monotonicity*: It is assumed that SWB questions are a positive monotonic transformation of the underlying concept called welfare/utility.
- *Ordinal comparability*: Using the welfare approach ordinal comparability is assumed, that is, if $SWB_i > SWB_j$, then $W_i > W_j$ for individual i and j. That is, individuals use discrete SWB rating in the same way and share a common opinion of what happiness is.
- *Interpersonal cardinality*: This assumption amounts to assuming that the difference between a satisfaction answer between 9 and 10 is the same as the difference between 2 and 3, and therefore we can make interpersonal comparisons.[6]

Monotonicity assumption implies that SWB actually measures welfare/utility. The problem is that welfare cannot be easily and objectively measured, and thus it is difficult to test some sort of correlation between both. Daniel Benjamin and colleagues have made contributions to this debate using experimental studies. For example Benjamin et al. (2012) analyze whether SWB questions are a good proxy for person's preferences, and hence a good proxy for utility. They conduct an experiment where respondents are faced to hypothetical scenarios and derive both choice and anticipated SWB[7] ranking of two alternatives. According to the results, authors argue that individuals do not seek to maximize SWB exclusively, but that SWB is a uniquely important argument of the utility function and thus it can be used as a proxy for utility. They also found that hypothetical choices maximize predicted life satisfaction (compare with other SWB questions) for most of the

[6]This can formally be stated as $(W_i - W_j) = \omega(SWB_i, SWB_j)$, with $\omega(\cdot)$ a function that is known up to a multiplicative constant. Normally $\omega(SWB_i, SWB_j)$ is taken to be $(SWB_i - SWB_j)$ (Ferrer-i Carbonell and Frijters, 2004).

[7]In order to obtain anticipated SWB they ask whether the respondent anticipates more life satisfaction for a particular hypothetical situation.

respondents in most of the scenarios. However, in settings where one alternative involves higher income or more money, they found that respondents are systematically more likely to choose the money alternative than they are likely to predict it will yield higher SWB. In other words, individuals are more sensitive to money in hypothetical choices than in anticipated SWB measures. In view of Equation (3.2), the implication of this result is that CVs obtained by any regression may be upward biased.

Benjamin et al. (2014) ask directly whether CVs (or marginal rate substitution) based on SWB reflect preference-based CVs. Using residential choices of medical students, they find substantial differences between the CVs implied by anticipated SWB measures and those revealed by choices. However, when searching which of the anticipated SWB matches most closely to the choices, and similar to Benjamin et al. (2012), they found that ladder-type questions and life satisfaction do better than happiness.

Both previous studies uses hypothetical well-being data, that is, they ask whether the respondent anticipates more life satisfaction for a particular hypothetical situation. In contrast, Perez-Truglia (2015) test the validity of experienced (as opposed to anticipated) well-being data. Perez-Truglia (2015) find that SWB is correlated with objective measures of well-being, such as suicide rates and frequency of smiling. But, life satisfaction performs significantly better than other objective and subjective measures of well-being. In line with Benjamin et al. (2014) and Benjamin et al. (2012), Perez-Truglia (2015)'s evidence suggests that life satisfaction offers some useful information about utility.[8] In general, measures of reported SWB are viewed as valid and reliable empirical proxy to individual's

[8]There are some other studies that have also found a strong and positive correlation between stated SWB measures and emotional expressions. See for example Deiner et al. (1999) or Ferrer-i Carbonell and Frijters (2004) and the references cited therein.

utility (Diener et al., 1995; Welsch and Kühling, 2009).

In the context of SWB studies, economists have presented different views to deal with the ordinal comparability and interpersonal cardinality assumption. Broadly, the ordinal comparability assumption tells us that all individuals that evaluate their well-being by say 4 have about the same satisfaction (i.e., they are in the same indifference curve). According to Van Praag (2007) and Ferrer-i Carbonell and Frijters (2004) this is not an unreasonable assumption for respondents who have about the same cultural and linguistic background. Since this paper focuses entirely on Chile, this assumption seems plausible for this study. Another important consequence of this assumption if valid is that we might be able to use ordered Probit and Logit models or any other model based on latent variable formulation.

The interpersonal cardinality is one of the strongest assumptions. However, I am interested in the CVs, which are invariant to any monotonic transformation of the SWB function and intrinsically ordinal, thus it is not necessary to assume cardinality (See Levinson, 2012; Frey et al., 2010; Welsch and Kühling, 2009, for similar assumption).

In sum, this approach has its own weakness since it makes strong assumptions about preferences in the sense that it compares stated well-being of different individuals. Therefore, the link between true well-being and subjective well-being is not without shortcomings. However, the set of assumptions of this approach are not stronger than those of other methods used to compute compensating variations (such as the hedonic or contingent evaluation approach), which have their own sets of strong assumptions (Levinson, 2012; Welsch and Kühling, 2009). Finally, in Levinson (2012)'s words *"...the easiest way to think about this methodology is that it uses respondents' stated happiness as a proxy for their utility, or as an* *observable manifestation of latent utility. As long as respondents with higher latent utility are more likely to say they are happier, this approach is consistent with a wide variety of discrete choice models in economics".*

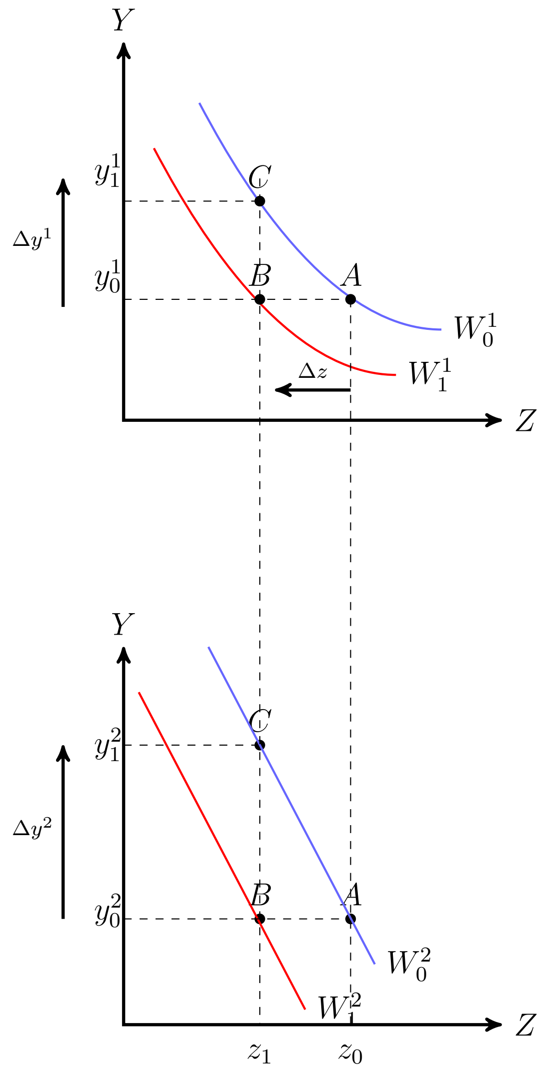
3.2.2 Spatial heterogeneity

Equations (3.1) and (3.2) assume fixed compensating variation for each attribute across space. However, one might expect that how people value attributes, and in particular neighborhood characteristics is not spatially stationary, but rather region-specific. As stated previously, the sensitivities measured by the marginal (dis)utilities of neighborhood attributes as well as the marginal utility of income may be different across the geographical space due to observed factors and unobserved factors. For example, how people respond to changes in Z may be different in urban and rural areas (observed spatial heterogeneity) or simply the preferences are intrinsically different across space (unobserved spatial heterogeneity).

This idea is shown in Figure 3.1. The top panel shows the indifference curve for a representative individual from region 1, while the bottom panel shows a representative individual from region 2, holding H and X being constant. People in region 2 are more sensitive to changes in neighborhood quality than those to region 1, which is represented by a steeper slope.

The representative individuals in both regions are initially located in point A with income y_0^1 and y_0^2, respectively, and the same amount of neighborhood quality z_0. Now assume that neighborhood quality decreases in both regions by $\Delta z = z_1 - z_0$, which implies a move to B where individuals from both are worse off with welfare W_1^1 and W_1^2, respectively. The compensating variation, which takes the

Figure 3.1: Compensating variation for two regions.



representative individual from both regions back to W_0^1 and W_0^2 at point C while accepting the worsening of neighborhood quality z_1, is represented $\Delta y = y_1 - y_0$ which is positive. However, individuals in region 2 are, in average, more sensitive to changes in neighborhood quality, that is, $(|\Delta y^2| > |\Delta y^1|)$, therefore they need a bigger CV, given the same change in Z. We might be able to remedy the foregoing concern by writing Equation (3.2) as

$$CV_c = -\frac{\left[\frac{\partial W}{\partial Z}\right]_c}{\left[\frac{\partial W}{\partial Y}\right]_c}, \tag{3.4}$$

where the marginal effects of Z and Y vary with location c. In other words, we can expect different CVs for each geographical location because individuals differentiate in terms of valuation of changes in neighborhood characteristics and/or in terms of different sensitivities to money changes among locations.

In Section 3.3, I present the econometric approach used in this work to model location-specific compensating variations. However, up to this point, it should be kept in mind some restrictions. First, both the theoretical model and subsequent econometric implementation allow for spatial heterogeneity across locations, but not within locations. However, this assumption is not as restrictive as assuming fixed sensitivities across geographical space on the underlying utility function. Second, due to restrictions in the estimation method further explained in Section 3.3.3, the calculation of compensating variations assumes that the marginal effect of income is fixed, that is, what makes CVs vary is just different sensitivities to the variable of interest.

3.3 Econometric modeling

This section presents the econometric modeling used in this study to implement the idea of spatial heterogeneous compensating variations . This is done in several steps. In Section 3.3.1, I discuss how to implement continuous and discrete spatial heterogeneity in the coefficients, and advantages and disadvantages of both approaches. Section 3.3.2 presents the estimation method. Finally, Section 3.3.3

shows how the location-specific compensating variations are computed.

3.3.1 Continuous and discrete spatial heterogeneity

I assume that Equation (3.1) for individual i in commune c can be operationalized by an underlying continuous but latent well-being process W^*_{ic} based on a linear combination of covariates given by:

$$\begin{aligned} W^*_{ic} &= \mathbf{x}'_{ic}\boldsymbol{\beta}_c + \upsilon_c + \epsilon_{ic} \\ \boldsymbol{\beta}_c &\sim g(\boldsymbol{\theta}) \end{aligned} \tag{3.5}$$

where $\mathbf{x}_{ic}$, for simplicity in the notation, groups all the variables (Y, H, X, Z); $\epsilon_{ic} \sim N(0, \sigma^2_\epsilon)$ is the error term, but since the scale of W^*_{ic} is not identified I normalize $\sigma_\epsilon = 1$; υ_c is a commune-specific unobserved factor which is independent of ϵ_{ic} and can be interpreted as capturing all those effects not captured by the variables at the commune level or as an unobservable random cluster component.[9]

A main feature in formulation (3.5) is that I assume that the marginal (dis)utility of predictors ($\boldsymbol{\beta}_c$) are not constant, but rather inherently different across geographical space. Spatial heterogeneity is modeled by allowing the parameters associated with each observed variable to vary randomly across space according to some distribution $g(\boldsymbol{\theta})$. However, we do not know how the parameters vary across space. All we know is that they vary locally with population probability density function (pdf) $g(\boldsymbol{\theta})$, which is assumed to be well-behaved. Since the distribution is unknown it has to be chosen a priori by the researcher. It is worth mentioning

[9]The normality assumption in the commune-specific unobserved factor is not required. However, this assumption is generally used in the applied literature to model random effect models (Greene, 2012; Greene and Hensher, 2010a).

that by assuming spatial random coefficients the CVs—which are the ratios of the estimated parameters— are also random and vary across space. In Section 3.3.3 I detail some caveats that must be considered when working with ratio of random parameters.

I compare and analyze two competitive ways of modeling spatial heterogeneity. First, I assume that there exists continuous spatial heterogeneity. This is introduced by assuming that $g(\boldsymbol{\theta})$ is a continuous function. This pdf can in principle take any shape, but the researcher must choose one according to his beliefs of the domain and boundedness of the coefficients. Once the distribution is chosen, 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 the distribution of $\boldsymbol{\beta}_c$. The goal is to estimate $\boldsymbol{\theta}$, and then get a profile of the distribution. As an illustration, we can assume that one of the coefficients is normally distributed. Therefore, we can write

$$\beta_c \sim g(\beta_c|\boldsymbol{\theta}) \implies \beta_c = \beta + \sigma_\beta \omega_c,$$

where $\omega_c \sim N(0,1)$, and $\boldsymbol{\theta} = (\beta, \sigma_\beta)'$. By the normality assumption, we make the implicit assumption that the coefficient in each commune can take any real number in the interval $(-\infty, +\infty)$. Obviously, any other distribution with some restrictions in the domain can be used.[10]

Instead of assuming a continuous distribution for the spatially random coefficients, one may also assume that they are distributed across space following a discrete distribution. In this case, spatial heterogeneity is modeled by assuming

[10] See for example Sarrias (2015a) for a discussion about some distributions and their implications.

that there is a discrete number, Q, of classes of communes where the communes belonging to the same class share the same coefficients. Formally, the distribution can be written as

$$\beta_c \sim g(\beta_q|\boldsymbol{\theta}_q) \implies \beta_c = \beta_q \quad \text{with probability} \quad w_{cq}, \tag{3.6}$$

where commune c belongs to class q with probability w_{cq}, such that $\sum_q w_{cq} = 1$. The researcher does not know from the sample which region is in which class, hence the assignment probability in (3.6) is unknown. Thus, the number of classes Q and the discrete distribution must be chosen a priori. The most widely used formulation for w_{cq} is the semi-parametric multinomial logit format

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

where $\mathbf{h}_c$ represents a vector of communes' 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. 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.

Before going further, it is important to discuss some advantages and disadvantages of both methods. As previously stated, one feature of the continuous case is that we need to specify the distribution of the coefficients a priori. If the distribution for the unobserved spatial heterogeneity is not properly selected, then a problem of misspecification might arise. On the other hand, the discrete approach is a semi-parametric specification. Thus, no assumptions about the shape of the spatial heterogeneity in terms of domain and boundedness are made. This frees the researcher from potential problems of misspecification. The only decision about the distribution is the number of classes—which is the same as the number support points. A disadvantage of the discrete case regarding to the last point is that since Q is not a parameter, hypotheses on Q cannot be tested directly.

In terms of the extent of unobserved spatial heterogeneity captured by both models, the discrete case is less flexible than the continuous case since the former assumes that the communes in the same class have the same parameters, while the latter allows for different sensitivities for each commune. Additionally, in the discrete approach we can observe a rapid increase in the number of parameters. For example, in a model with 10 variables and 4 classes we will have ($10 \times 4 + 3$ =) 43 parameters. Furthermore, it is often observed that some parameters collapse to the same value across classes or some classes obtaining very small probabilities, which is highly likely in the case of strongly peaked distributions (Hess, 2014). However, as illustrated in this study, the continuous approach is not immune to the proliferation-of-parameters problem if a correlated random parameter model is specified and the number of variables is high.

The discrete approach has some advantages over the continuous approach in terms of estimation and computational cost. As I will discuss in Section 3.3.2, the discrete case does not require simulation-based methods for the estimation of the parameters, while the continuous case requires the simulation of the integration over all the possible values of $\boldsymbol{\beta}_c$ for the estimation. As I will discuss later, this is one of the main issued to be considered when comparing both models.

3.3.2 Estimation

Since the observed SWB measure is binary, the latent variable W_{ic}^* is linked to the observed binary variable W_{ic} by the following rule

$$W_{ic} = \begin{cases} 1 & \text{if } W_{ic}^* > 0 \\ 0 & \text{if } W_{ic}^* \leq 0 \end{cases}.$$

Then, the probability for individual i in commune c, given the latent structure in Equation (3.5) is:

$$\Pr(W_{ic}^* = W_{ic}|\mathbf{x}_{ic}, \boldsymbol{\beta}_c) = \left[\Phi\left(\mathbf{x}_{ic}'\boldsymbol{\beta}_c\right)\right]^{W_{ic}}\left[1 - \Phi\left(\mathbf{x}_{ic}'\boldsymbol{\beta}_c\right)\right]^{1-W_{ic}},$$

where $\Phi(\cdot)$ denotes the cumulative density function of the standard normal distribution. If there are C different communes with n_c individuals in each of them, and assuming that individuals are independent across communes, the joint probability density function, given $\boldsymbol{\beta}_c$ for commune c is:

$$\Pr(\mathbf{w}_c|\mathbf{X}_c, \boldsymbol{\beta}_c, \upsilon_c) = \prod_{i=1}^{n_c} \Pr(W_{ic}^* = W_{ic}|\mathbf{x}_{ic}, \boldsymbol{\beta}_c), \tag{3.7}$$

where $\mathbf{w}_c$ is the sequence of choices for all individuals in commune c. Since $\boldsymbol{\beta}_c$ is common for individuals living in the commune c, within each commune individuals are not independent. Thus, the unconditional pdf of $\mathbf{w}_c$ given $\mathbf{X}_c$ will be the weighted average of the conditional probability (3.7) 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

$$
\begin{aligned}
P_c(\boldsymbol{\theta}_q) &= \sum_{q=1}^{Q} w_{cq} \left[\prod_{i=1}^{n_c} \Pr(W_{ic}^* = W_{ic} | \mathbf{x}_{ic}, \boldsymbol{\beta}_c, \boldsymbol{\theta}_q) \right], && (3.8) \\
P_c(\boldsymbol{\theta}) &= \int_{\boldsymbol{\beta}_c} \left[\prod_{i=1}^{n_c} \Pr(W_{ic}^* = W_{ic} | \mathbf{x}_{ic}, \boldsymbol{\beta}_c, \boldsymbol{\theta}) \right] g(\boldsymbol{\beta}_c) d\boldsymbol{\beta}_c. && (3.9)
\end{aligned}
$$

The discrete spatial heterogeneity model with unconditional pdf given by (3.8) can be estimated using maximum likelihood approach. However, the unconditional pdf of the continuous case given by (3.9) has no closed form solution, so the log-likelihood function is hard to compute. To overcome this problem, we can simulate this probability and use the simulated maximum likelihood (SML) approach to estimate $\boldsymbol{\theta}$ (Gourieroux and Monfort, 1997; Hajivassiliou and Ruud, 1986; Stern, 1997; Train, 2009). 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 (3.9) is calculated. This process is repeated for many draws, and the average over the draws is the simulated probability. Formally, the simulated probability for commune c is

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

where $\widetilde{P}_{icr}$ is the probability for individual i in commune 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:

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

Lee (1992), Gourieroux and Monfort (1991) and Hajivassiliou and Ruud (1986) derive the asymptotic distribution of the SML estimator based on smooth probability simulators with the number of draws increasing with sample size.[11]

3.3.3 Location-specific compensating variations

In order to compute commune-specific compensating variations I need the ratio between the coefficient of the variable of interest and the coefficient of income. However, some difficulties arise if we assume that both coefficients are random. To show this point, consider expanding the latent welfare function (3.5) as a linear form of income and variables $\mathbf{x}_{ci}$,

$$W^*_{ci} = \alpha_c \text{income} + \mathbf{x}'_{ci}\boldsymbol{\beta}_c + \epsilon_{ic}$$

Then, according to Equation (3.4), CV for variable k in commune c is given by:

$$CV_{ck} = -\frac{\beta_{kc}}{\alpha_c} \tag{3.10}$$

where $\beta_{kc} \sim g(\beta_k, \sigma_{\beta_k})$ is the coefficient of the variable of interest for individual i in commune c, and α_c is the coefficient for income, which also varies across individuals according to $\alpha_c \sim h(\alpha, \sigma_\alpha)$.

The main shortcoming with the CV given in Equation (3.10), when both distributions are continuous, is that the ratio of two estimated distributions may

[11] All the estimations conducted in this study are carried out using R software Team (2015). In particular, I used the Rchoice package (Sarrias, 2015b) to estimate the models using simulated maximum likelihood. The ML algorithm for the discrete case was also coded in the same software.

not result in a well-specified distribution. For example, the ratio of two normal distributions produces a Cauchy distribution with no finite moments (see for example Daly et al., 2012, for other examples).[12] In order to avoid this problem, I assume that the marginal utility of household income, α, is fixed across communes when estimating a continuous spatial heterogeneous model. This allows the distribution for the kth compensating variation to be the same as the distribution of the kth variable scaled by α. For example, if $\beta_{ck} \sim \mathrm{N}(\beta_k, \sigma_k^2)$, then $(1/\alpha)\beta_{ck} \equiv \mathrm{N}(\beta_k/\alpha, \sigma_k^2/\alpha)$.

To compute the most likely location of a given commune's CV on the population distribution we move from the unconditional to the conditional distribution using Bayes' theorem. Formally, if $h(\boldsymbol{\beta}_c)$ is a function of $\boldsymbol{\beta}_c$ (as the CVs in Equation (3.10)), then

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

where $f(h(\boldsymbol{\beta}_c)|\mathbf{w}_c, \mathbf{X}_c, \boldsymbol{\theta})$ is the distribution of $h(\boldsymbol{\beta}_c)$ conditional on the observed choices of individuals in commune c, and $g(\boldsymbol{\beta}_c|\boldsymbol{\theta})$ is the unconditional distribution. The conditional expectation of $h(\boldsymbol{\beta}_c)$ is thus given by:

$$\mathbb{E}\left[h(\boldsymbol{\beta}_c)|\mathbf{w}_c, \mathbf{X}_c, \boldsymbol{\theta}\right] = \frac{\int_{\boldsymbol{\beta}_c} h(\boldsymbol{\beta}_c)f(\mathbf{w}_c|\mathbf{X}_c, \boldsymbol{\beta}_c)g(\boldsymbol{\beta}_c|\boldsymbol{\theta})d\boldsymbol{\beta}_c}{\int_{\boldsymbol{\beta}_c} f(\mathbf{w}_c|\mathbf{X}_c, \boldsymbol{\beta}_c)g(\boldsymbol{\beta}_c|\boldsymbol{\theta})d\boldsymbol{\beta}_c}. \tag{3.11}$$

The expectation in Equation (3.11) gives us the conditional mean of the distribution of the random parameters, which can also be interpreted as the posterior distribution of the individual parameters. Simulators for this conditional expecta-

[12] When the distribution is discrete, and all the points for the income coefficient are estimated to be away from zero, the implied CV distribution is also finite with defined values (Daly et al., 2012).

tion are presented below for the continuous and discrete, respectively:

$$\bar{h}(\widehat{\boldsymbol{\beta}}_c) = \widehat{\mathbb{E}}\left[h(\boldsymbol{\beta}_c)|\mathbf{w}_c, \mathbf{X}_c, \boldsymbol{\theta}\right] = \frac{\frac{1}{R}\sum_{r=1}^{R} h(\widehat{\boldsymbol{\beta}}_{cr})\prod_i f(W_{ci}|\mathbf{x}_{ci}, \widehat{\boldsymbol{\beta}}_{cr}, \widehat{\boldsymbol{\theta}})}{\frac{1}{R}\sum_{r=1}^{R}\prod_i f(W_{ci}|\mathbf{x}_{ci}, \widehat{\boldsymbol{\beta}}_{ci}, \widehat{\boldsymbol{\theta}})} \tag{3.12}$$

$$\bar{h}(\widehat{\boldsymbol{\beta}}_c) = \widehat{\mathbb{E}}\left[h(\boldsymbol{\beta}_c)|\mathbf{w}_c, \mathbf{X}_c, \boldsymbol{\theta}_q\right] = \frac{\sum_{q=1}^{Q} h(\widehat{\boldsymbol{\beta}}_q)\widehat{w}_{cq}\prod_i f(W_{ci}|\mathbf{x}_{ci}, \widehat{\boldsymbol{\beta}}_{cq}, \widehat{\boldsymbol{\theta}}_q)}{\sum_{q=1}^{Q}\widehat{w}_{cq}\prod_i f(W_{ci}|\mathbf{x}_{ci}, \widehat{\boldsymbol{\beta}}_{ci}, \widehat{\boldsymbol{\theta}}_q)} \tag{3.13}$$

In order to construct the confidence interval for $\bar{h}(\widehat{\boldsymbol{\beta}}_c)$, we can derive an estimator of the conditional variance from the point estimates as follows (Greene, 2012, chap. 15):

$$\widehat{V}_c = \widehat{\mathbb{E}}\left[h(\boldsymbol{\beta}_c)^2|\mathbf{y}_c, \mathbf{X}_c, \boldsymbol{\theta}\right] - \widehat{\mathbb{E}}\left[h(\boldsymbol{\beta}_c)|\mathbf{y}_c, \mathbf{X}_c, \boldsymbol{\theta}\right]^2. \tag{3.14}$$

An approximate normal-based 95% confidence interval can be then constructed as $\bar{h}(\widehat{\boldsymbol{\beta}}_c) \pm 1.96 \times \widehat{V}_c^{1/2}$.

3.4 Data and results

This Section presents the empirical application using a Chilean dataset that illustrates the differences in the econometric approach presented in Section 3.3. I first look at data before discussing the main estimation results for the fixed, continuous and discrete case. Finally, I broadly discuss the results for the compensating variations distributions obtained using both methods.

3.4.1 Data

I use data from the 2013 National Socioeconomic Characterization Survey (CASEN) from Chile.[13] CASEN is a national population based survey, which is representative at the communal level and is carried out by the Ministry of Planning (MIDEPLAN) to describe the socioeconomic situation as well as the impact of social programs on the living conditions of the Chilean Population. Our sample in this study corresponds to 16,008 individuals between 15 and 65 years old in 324 communes after cleaning by missing values in the covariates.[14]

The dependent variable in this study is obtained from the response to the question "*How satisfied are you with life?*". The answer ranges from 1 "Completely unsatisfied " to 10 "Completely satisfied". As I reviewed in Section 3.2.1, this SWB question is considered as a better proxy of welfare, compare to other SWB questions. For sake of simplicity of interpretation and estimation, the response is recoded as a binary variable. Namely, the respondents are classified as being "Satisfied" if they replied 8, 9, or 10, and being "Unsatisfied or neither satisfied nor unsatisfied" if they replied with an integer between 1 and 7. Table 3.1, which reports summary statistics, shows that almost half of the sample report being satisfied with life.[15]

I control for several individual's characteristics (X) that are common in the literature of SWB. Table 3.1 shows that in our sample the average schooling is $\approx$ 10 years, and males account for almost 33% of the total sample. Unemployed individuals account just for a small part (5%) of the sample, whereas individuals

[13]See `http://observatorio.ministeriodesarrollosocial.gob.cl/casen/casen_obj.php`.

[14]Chile has 346 communes of which 324 are representative in CASEN 2013.

[15]The estimations were also carried out using other dichotomizations, however no significant differences were observed. The results are available upon request.

Table 3.1: Summary statistics.

Statistic	N	Mean	St. Dev.	Min	Max
Satisfied with life	16,008	0.528	0.499	0	1
Log(Household Income pc)	16,008	12.009	0.785	0.000	16.611
Schooling	16,008	9.967	3.980	0	22
Age	16,008	42.564	13.476	15	65
Male	16,008	0.330	0.470	0	1
Unemployed	16,008	0.043	0.202	0	1
OLF	16,008	0.409	0.492	0	1
Married	16,008	0.601	0.490	0	1
Household size	16,008	3.610	1.618	1	17
Disability	16,008	0.067	0.250	0	1
Treatment	16,008	0.301	0.459	0	1
Accident	16,008	0.214	0.410	0	1
Noise pollution	16,008	0.163	0.369	0	1
Air pollution	16,008	0.167	0.373	0	1
Water pollution	16,008	0.073	0.260	0	1
Garbage pollution	16,008	0.142	0.349	0	1
Robbery	16,008	0.309	0.462	0	1
Poor surveillance	16,008	0.316	0.465	0	1
Street drugs & alcohol	16,008	0.280	0.449	0	1
Drug trafficking	16,008	0.145	0.352	0	1
Log(Density)	16,008	3.813	2.354	−1.514	9.558
Log(Median income)	16,008	12.302	0.264	11.667	13.816
Urban	16,008	0.670	0.470	0	1

Note: Individual covariates and variables based on income come from the 2013 National Socioeconomic Characterization Survey (CASEN) from Chile. Density comes from the Chilean National Institute of Statistic (INE).

out of the labor force account for almost a 40%. In terms of marital status, 60% of the sample corresponds to people that are married or have a domestic partnership. Y is approximated by the household income per capita. The mean of household income per person is about $\exp(11.521) \approx \$164,226$ Chilean pesos.[16]

The component H in Equation (3.1) is captured by three variables. *Disability* equals 1 if the respondent has some kind of long-term disability, such as blindness,

[16]This is equivalent to US$235.

deafness, and so on. The variable *treatment* indicates whether the individual is currently taking some medical treatment for the following illness: hypertension, diabetes, depression, heart attack or cancer. Finally, *accident* equals 1 if the individual have had an accident or sickness in the last 3 months. Table 3.1 shows that 6.7% of the sample has some long-term disability, while 30% is taking some type of treatment and 21% have had some kind of accident or sickness.

In addition to information about individuals characteristics and health status, I use a set of dummies variables indicating neighborhood attributes, denoted by Z in Equation 1. Individuals are asked about their neighborhood experience defined by the perception of different problems and characteristics within their neighborhood or location. The first of dummy are related to self-perception of pollution and environmental problems. Specifically, the dummies variables are obtained from the response to the question: *"What problems related to pollution and environmental degradation do you identify in your neighborhood or location?"* Based on the answer to this question different dummy variables were created for the following problems: 1) noise pollution, 2) air pollution, 3) water pollution, and 4) garbage in the neighborhood. The proportion of individuals reporting these problems are 16%, 17%, 0.7% and 21% respectively.

Similarly, individuals were asked *"What problems related to public security do you identify in your neighborhood or location?"* Based in the answer to this question, several dummies variables were created for the following security problems: 1) robberies and assaults on people, homes and/or vehicles (*Robbery*), 2) drug traffic (*Drug traffic*), 3) outbreaks of alcohol or drugs on the street (*Street drugs & alcohol*), and 4) insufficient police surveillance (*Poor surveillance*).

Finally, the following controls at the commune level are used: logarithm of median income as a proxy for commune's development, logarithm of density and a dummy variable indicating if the individual is currently living in a urban area.

3.4.2 Fixed coefficients

In this section I comment on the results of the model with fixed parameters (i.e. not assuming spatial heterogeneity and our benchmark model), which is presented in the first three columns of Table 3.2. The first column gives the estimated coefficients, while the second column and third column display the standard errors and the compensating variation, respectively. While the magnitude of the estimated coefficients does not have a direct interpretation for binary Probit models, their sign indicates the direction of the relationship. For example, a positive coefficient means a positive correlation on the latent SWB, thus a higher probability of reporting good satisfaction with life.

As expected by theoretical and empirical studies, the estimated parameters indicate a significant positive relationship between household income pc[17] and well-being; in other words people with higher household income are more likely to report higher subjective well-being. Similarly, more educated individuals are more satisfied with their life. Apart from schooling and income, it can be observed that men have, on average, higher levels of life satisfaction than women, whereas unemployed, and older people are more likely to be less satisfied with their lives. In line with the previous literature, I also found that being out of labor force (compared to being employed), living with a partner and bigger household size also increases the probability that a person will report high well-being. All variables related to negative health status have the expected negative sign, and are highly signifi-

[17]Due to frequent use, 'per capita' is abbreviated as 'pc'.

Table 3.2: Continuous spatial heterogeneity.

	Fixed Parameters			**All Normals**				**Correlated**				**Correlated (LN & CN)**				
	Est.	*SE*	*CV*	*Est.*	*SE*	*CV*	$\% > 0$	*Est.*	*SE*	*CV*	$\% > 0$	*Est.*	*SE*	*Mean*	*SD*	*CV*
A: Mean Parameters																
Log(Hincome)	0.236***	0.016		0.239***	0.017			0.265***	0.019			0.260***	0.012			
Constant	−2.970***	0.532		−2.970***	0.708			−3.186***	0.978			−3.122***	0.617			
Schooling	0.032***	0.003	-0.137	0.033***	0.003	-0.140	100%	0.032***	0.005	-0.120	88%	−3.501***	0.125	0.034	0.017	-0.130
Age	−0.005***	0.001	0.020	−0.005***	0.001	0.022	3%	−0.006***	0.001	0.024	23%	−0.005***	0.001			0.021
Male	0.045*	0.024	-0.189	0.043*	0.024	-0.178	69%	0.052*	0.027	-0.196	69%	0.037	0.025			-0.144
Unemployed	−0.217***	0.052	0.917	−0.217***	0.053	0.905	2%	−0.214***	0.061	0.808	22%	−2.012***	0.531	0.225	0.304	-0.866
OLF	0.082***	0.024	-0.346	0.084***	0.024	-0.352	84%	0.085***	0.027	-0.321	77%	0.088***	0.025			-0.337
Married	0.194***	0.022	-0.820	0.195***	0.023	-0.817	99%	0.221***	0.025	-0.835	89%	0.207***	0.023			-0.798
Household Size	0.030***	0.007	-0.127	0.029***	0.007	-0.120	100%	0.034***	0.009	-0.127	68%	0.036***	0.007			-0.137
Disability	−0.225***	0.042	0.952	−0.226***	0.044	0.944	1%	−0.245***	0.051	0.925	17%	−1.699***	0.345	0.223	0.155	-0.857
Treatment	−0.152***	0.025	0.643	−0.154***	0.026	0.645	4%	−0.155***	0.029	0.584	18%	−2.397***	0.319	0.133	0.143	-0.514
Accident	−0.070***	0.026	0.298	−0.069**	0.027	0.287	23%	−0.068*	0.042	0.258	37%	−3.156***	0.694	0.085	0.145	-0.326
Noise pollution	−0.076***	0.029	0.321	−0.070**	0.030	0.294	23%	−0.075	0.054	0.282	40%	−0.228	0.320			0.878
Air pollution	−0.014	0.028	0.060	−0.009	0.030	0.039	46%	0.003	0.033	-0.012	51%	−0.181	0.732			0.696
Water pollution	−0.044	0.040	0.187	−0.044	0.042	0.184	33%	−0.024	0.055	0.091	47%	−0.720	1.192			2.772
Garage pollution	−0.003	0.030	0.012	−0.003	0.032	0.013	49%	−0.017	0.041	0.066	48%	−1.050	0.816			4.043
Robbery	−0.042*	0.023	0.179	−0.042*	0.025	0.177	33%	−0.041	0.036	0.154	43%	−3.820***	1.039	0.053	0.115	-0.203
Poor surveillance	−0.053**	0.023	0.225	−0.052**	0.024	0.218	27%	−0.038	0.034	0.145	43%	−4.056***	1.089	0.054	0.156	-0.206
Street drugs	−0.056**	0.025	0.236	−0.051*	0.027	0.213	30%	−0.062	0.051	0.234	41%	−3.823***	1.140	0.058	0.140	-0.222
Drug trafficking	−0.071**	0.032	0.301	−0.070**	0.034	0.293	23%	−0.079	0.041	0.298	42%	−3.714***	1.133	0.087	0.295	-0.334
Log(Density)	−0.007	0.005	0.031	−0.010	0.006	0.041	0%	−0.009	0.008	0.034	40%	−0.011*	0.006			0.043
Log(Income)	−0.004	0.046	0.016	−0.004	0.060	0.017	18%	−0.012	0.080	0.045	40%	−0.014	0.052			0.054
Urban	0.078***	0.024	-0.328	0.074***	0.025	-0.309	83%	0.082***	0.031	-0.308	60%	0.068**	0.026			-0.261
B: Standard Deviations																
υ_c				0.065**	0.028			0.866	1.373			1.631	1.305			
Schooling				0.001	0.002			0.027***	0.005			0.481***	0.118			
Age				0.003***	0.000			0.009***	0.002			0.008***	0.002			
Male				0.086**	0.038			0.107**	0.042			0.059	0.050			
Unemployed				0.101	0.133			0.277***	0.099			1.020**	0.433			
OLF				0.085**	0.039			0.118**	0.050			0.118***	0.044			
Married				0.078***	0.028			0.178***	0.035			0.175***	0.046			
Household Size				0.011	0.008			0.072***	0.009			0.061***	0.013			
Disability				0.096	0.094			0.256***	0.068			0.627*	0.357			
Treatment				0.087**	0.038			0.169***	0.038			0.875***	0.205			
Accident				0.092*	0.053			0.201***	0.051			1.171***	0.437			
Noise pollution				0.097	0.062			0.304***	0.055			0.478*	0.258			
Air pollution				0.089**	0.045			0.145***	0.041			0.204	0.496			
Water pollution				0.098	0.161			0.361***	0.086			0.789	0.772			
Trash Poll.				0.106	0.078			0.320***	0.065			0.829*	0.457			
Robbery				0.095**	0.040			0.227***	0.059			1.323**	0.553			
Poor surveillance				0.086*	0.045			0.210***	0.036			1.502***	0.524			
Street drugs				0.098	0.068			0.287***	0.040			1.391**	0.610			
Drug trafficking				0.097	0.092			0.384***	0.063			1.593**	0.626			
Log(Density)				0.002	0.005			0.035***	0.010			0.027**	0.012			
Log(Income)				0.004*	0.002			0.045	0.061			0.117	0.100			
Urban				0.079**	0.035			0.339***	0.038			0.305***	0.038			
LL	-10560			-10500				-10270				-10330				
# parameters	23			45				276				276				
n	16008			16008				16008				16008				
AIC	21176			21081				21100				21206				
BIC	21352			21426				23220				23326				

Note: Simulation based on 100 Halton draws. Standard errors in parentheses. $AIC = -2\log L + 2K$, $BIC = -2\log L + \log(N)K$. $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.001$

cant: having an accident, some kind of disability or medical treatment increases the probability of reporting lower levels life satisfaction. Table 3.2 also shows that noise pollution in the neighborhood correlates negatively with life satisfaction and is strongly significant. Nevertheless, the rest of the variables measuring perceived pollution in the neighborhood (air, water and garbage pollution), even though negative, are not significant. The results also show a negative and statistically significant relationship between neighborhood problems and individual life satisfaction. For example, robbery, poor surveillance, street alcohol and drug trafficking problems in the neighborhood are negative and statistically different from zero.

The only variable at the commune level that enters significantly is urban. The estimated coefficient implies that individuals living in urban areas are on average more satisfied with their life than those living in rural areas. In general, the sign of the coefficients are consistent with the applied literature. The only intricate result is that for the log of the median income at the commune level. The insignificant coefficient implies that the level of development of the commune does not have an effect on individuals' well-being.

However, these coefficients should not be overemphasized, since they require the assumption of interpersonal cardinality, which is a strong assumption. Given the above, it is more appropriate and conservative to analyze compensatory variations.

Compensating variations for the spatially stationary model are presented in column 3. The figures are computed as minus the ratio between the coefficient for the variable in interest and the coefficient for the logarithm of household income per capita taken from column 1 in Table 3.2. It is important to stress that when changes in the variables indicate a worsening (improvement) of well-being, individuals require an increase (decrease) in household income pc to compensate for this loss (gain), then we should observe a positive (negative) sign for the ratio. For example, an individual with perceived noise pollution in the neighborhood would require, in average, a relative increase of 0.321 (or an increase of 32.1%) in his household income pc in order to remain in the respective satisfaction level with the externality.[18] An important observation is that a disability situation is the detrimental effect that requires the highest compensation: a disabled person requires a relative increase of income of about .95. This is followed by unemployment with a relative increase of .92. This last result is consistent with the finding from previous studies on the non-pecuniary cost of unemployment on an individual's subjective well-being (see Winkelmann and Winkelmann, 1998; Clark and Oswald, 2002; Powdthavee, 2008; Mentzakis, 2011). Among the perceived neighborhood problems, drug trafficking is the situation that requires the highest compensation. For example, on average, an individual living in a neighborhood with drug trafficking would require an increase of about 30% of the household income pc in order to be unaffected by this problem. How much do people value living in urban areas? The results show that the benefit of moving from rural to urban communes is worth about as a relative increase of $\Delta Y/Y_0 = .33$. Strictly speaking, an individual living in a rural area with income $Y_0 + \Delta Y/Y_0$ would have the same level of well-being as an individual living in a urban area with income Y_0. Thus, $\Delta Y/Y_0$ can be considered as the shadow price of urbanism.

3.4.3 Continuous spatial heterogeneity

The coefficients in the previous section reflect the average marginal impact of the variables and their respective CVs for some average individual in the sample. However, they obscure the potential latent heterogeneity across communes. The second model of Table 3.2 shows the results for a model where all the parameters, except for the logarithm of household income pc, are assumed to be independently normally distributed across space. That is, I assume that the coefficients vary

[18] Since household income pc is in log, we can interpret CVs as semi-elasticities.

randomly across communes and therefore they can take positive or negative values.[19] As explained before, the household income pc is held fixed to prevent the non-existence of moments for the ratio when computing the CVs. The coefficient vector is expressed as $\boldsymbol{\beta}_c = \boldsymbol{\beta} + \mathbf{L}\boldsymbol{\omega}_c$, where $\mathbf{L}$ is a diagonal matrix whose elements are standard deviation and $\boldsymbol{\omega} \sim \mathrm{N}(0, 1)$. For the simulation procedure 100 Halton draws were used for each commune and parameter in each specification.[20] For each parameter, the mean and standard deviation of the distribution is estimated. Subsequent models allow correlation among the coefficients and specify log-normal and truncated normal distribution for some of the coefficients.

The results are broadly robust by allowing continuous spatial heterogeneity: both the mean coefficients and the CVs are very close in magnitude, vis-a-vis, to those for the benchmark (stationary) model. The standard deviations for some of the variables are highly significant, indicating that those parameters do indeed vary across communes, and focusing on the central tendency alone veils useful information. I conducted a LR test comparing model 1 and model 2. The statistic is $\chi^2_{22} = 139.04$, therefore we reject the null hypothesis that the standard deviations are jointly zero. The AIC criteria decreases, but the BIC criteria, which penalized for the number of parameters, increases.

The parameters with significant standard deviation—which measure the degree of spatial heterogeneity—are age, male, inactive, married, medical treatment, ac-

[19] I have also tried other distributions in the specification such as the triangular, uniform and Johnson S_b distribution, and other different combinations. However, models showed comparative lower fit than the models presented in Table 3.2.

[20] Good performance of SML requires a very large number of draws. However, the maximization of SML can be very time consuming with large and complex models. Researchers have gained speed with no degradation in simulation performance through the use of smaller number of Halton draws (Bhat, 2001; Train, 2000). Bhat (2001)'s Monte Carlo analysis found that the precision of the estimated parameters was smaller using 100 Halton draws than 1000 pseudo-random number in the context of Mixed Logit. In this study, I found that beyond 100 Halton draws did not lead to significant changes in the estimated parameters.

cident, air pollution, robbery, poor surveillance, log of communal median income and urban. Since the parameters are normally distributed we can get an approximate proportion of communes with positive and negative coefficients (see column 7).[21] For example, the mean and standard deviation of the schooling parameter implies that almost 100% of the communes have a positive correlation between schooling and well-being. Furthermore, the standard deviation of this parameter is not significant, implying that it is highly probable that coefficient does not vary across space, but rather is fixed and unique for all the communes. We explore this option later. The standard deviation of the commune-specific unobserved factor υ, which is assumed to be normally distributed is significant. This unobserved factor is intended to capture all those effects not captured by the variables at the commune level included in the model.

In the previous model, the coefficients are specified to be independently distributed. However, we might expect some degree of correlation between the coefficients. To allow this feature, I specify $\boldsymbol{\beta}_c = \mathrm{N}(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ for a general $\boldsymbol{\Sigma}$. The matrix $\mathbf{L}$ is now a lower-triangular Chokeski factor of $\boldsymbol{\Sigma}$, such that $\mathbf{LL}' = \boldsymbol{\Sigma}$. In extensive form, the coefficient vector can be expressed as:

$$\begin{pmatrix} \beta_{1r} \\ \beta_{2r} \\ \vdots \\ \beta_{Kr} \end{pmatrix} = \begin{pmatrix} \beta_{1} \\ \beta_{2} \\ \vdots \\ \beta_{K} \end{pmatrix} + \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{K1} & a_{K2} & \dots & a_{KK} \end{pmatrix} \begin{pmatrix} \omega_{1r} \\ \omega_{2r} \\ \vdots \\ \omega_{Kr} \end{pmatrix}.$$

Thus, each single coefficient can be written as:

[21]The proportion of communes with a positive coefficient for some attribute is given by $\Phi(\widehat{\beta}/\widehat{\sigma}_\beta)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

$$\beta_{kr} = \beta_k + \sum_{h=1}^{K} a_{kh}\omega_{hr}.$$

The third model in Table 3.2 presents the results for the correlated spatially random parameter model. The standard deviations of the parameters are computed by diag $\widehat{\boldsymbol{\Sigma}}^{1/2}$, and the standard errors are computed using Delta Method. In terms of relative quality of the correlated parameter models for the given data, it can be observed that both measures of information–AIC and BIC–are higher compared with the model where all parameters are assumed to be independently normally distributed. In other words, the model that assumes independent spatial heterogeneous parameters is favored by both criteria. An important result is that all the standard deviations of the parameters estimates—except for υ_c and log of comunal median income—are highly significant, but most of the mean parameters for the neighborhood characteristics are no longer significant.

One important limitation of the previous models is that it is assumed that the coefficients can take positive and negative values in the communes, whether this is true or not. For example, according to the results for the correlated parameter model for 20% of the communes increasing individuals' education reduces satisfaction with life. However, it is not clear whether this is in fact true or is simply an artifact of the model specification (normal assumption) or data quality. Similarly, for some variables such as health problems, or problems in the neighborhood, it is reasonable to expect a negative sign for all the communes—but different intensities. The fourth model in Table 3.2 abandons this possibility for some of the variables. In particular, I assume that schooling, unemployed, disability, medical treatment, accident, robbery, poor surveillance, street drugs, and drug trafficking follow a log-normal distribution. Therefore, the coefficient for each of these variables can be written 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 mean and standard deviation of β_c are $\exp(\beta + \sigma_\beta^2/2)$ and mean $\times \sqrt{\exp(\sigma_\beta^2) - 1}$, respectively.

Note that the log-normal distribution has its domain in the positive real line $(0, \infty)$. For variables where a negative sign is expected I include the negative of the variable. This allows the coefficients to be negative without imposing a sign change in the estimation procedure. Schooling enters directly, such that all communes's coefficients are positive, and for the rest of the variables assumed to be log-normal distributed are entered as negative.

For the pollution variables, I assume a truncated normal distribution, whose domain is also $(0, \infty)$. The parameter for each region is created as $\beta_c = \max(0, \beta + \sigma_\beta \eta_c)$ where $\eta_c \sim \mathrm{N}(0,1)$. The share of communes massed at zero—i.e. share of communes for which the relationship is assumed to be null—is equal to $\Phi(-\beta/\sigma_\beta)$.

One result worth mentioning in this last model is that the mean parameters of the variables related to problems in the neighborhood are now statistically significant. In terms of commune-level variables, we can see that the mean of density is negative and statistically significant. Finally, both AIC and BIC indicate that the fit is not good enough compared to previous models.

3.4.4 Discrete spatial heterogeneity

Table 3.3 presents two models that assume that the coefficients follow a discrete distribution across space. The first model assumes that there exist 2 classes, whereas the second assumes three classes. In both models, I assume that the probability of belonging to some class depends on the geographical location of each commune represented by the geographical coordinates of the centroids x and y, which represent the longitude (east and west) and latitude (north and south) respectively.[22]

A first important result is that both models present a better statistical fit compared to the reference model with fixed coefficients and all the models for continuous spatial heterogeneous models, while the model with 2 classes presents a weak superiority over the model with 3 classes by presenting a lower BIC. In other words, the model with three classes of communes is not able to retrieve significant amount of additional spatial heterogeneity when compared with the model with two classes. Given this, I will comment on the results of the model with 2 classes.

The estimated coefficients of the variables for the class assignment indicate that communes located further south and east are more likely to be in the second class. Therefore, given Chile's territorial shape the first class can be though as composed by communes located further north and closer to the coast, whereas the second class is characterized by communes located further south and closer to the Andes.

In general, the following can be noted from the model:

- Individual characteristics (X_i): Education is more valued in communes belonging to the second class. For example, a reduction of one year of education requires an increase of household income pc of approximately 15% (in average) compared to 1% of the first class, ceteris paribus. Both are statistically significant. Age is also valued negatively in both classes, however the compensation is higher in the first class (2.7% vs. 1.5%). Regarding to gender

[22] I have also estimated versions of the continuous model where the mean of the random parameters varies according to functions of the geographical coordinates. However, due to the complexity of the estimation procedure, the models in some cases did not converge or I encountered flat regions of the simulated likelihood function producing a singular Hessian.

Table 3.3: Discrete spatial heterogeneity.

	Model 1						**Model 2**								
	Class 1			***Class 2***			***Class 1***			***Class 2***			***Class 3***		
	Est.	*SE*	*CV*	*Est.*	*SE*	*CV*	*Est.*	*SE*	*CV*	*Est.*	*SE*	*CV*	*Est.*	*SE*	*CV*
Log(Household Income pc)	0.365***	0.037		0.196***	0.014		0.365***	0.039		0.177***	0.017		0.266***	0.040	
Constant	−7.385***	1.317		−1.187*	0.701		−7.435***	1.407		−1.602*	0.903		−0.791	1.607	
Schooling	0.035***	0.007	-0.096	0.029***	0.004	-0.149	0.033***	0.007	-0.090	0.053***	0.006	-0.299	−0.003	0.008	0.013
Age	−0.010***	0.002	0.027	−0.003**	0.001	0.015	−0.011***	0.002	0.031	0.002	0.002	-0.011	−0.010***	0.002	0.038
Male	−0.021	0.050	0.058	0.068**	0.031	-0.345	−0.040	0.053	0.109	0.070	0.045	-0.398	0.064	0.057	-0.241
Unemployed	−0.322***	0.112	0.883	−0.155**	0.067	0.791	−0.370***	0.119	1.015	−0.062	0.099	0.350	−0.268**	0.120	1.006
OLF	0.013	0.050	-0.036	0.119***	0.031	-0.606	0.007	0.053	-0.019	0.115**	0.045	-0.652	0.121**	0.057	-0.453
Married	0.144***	0.047	-0.395	0.225***	0.029	**-1.150**	0.158***	0.049	-0.435	0.283***	0.042	-1.601	0.121**	0.053	-0.454
Household Size	0.043***	0.016	-0.117	0.024***	0.009	-0.123	0.027*	0.016	-0.074	0.060***	0.013	-0.339	−0.010	0.016	0.038
Disability	−0.250***	0.087	0.685	−0.198***	0.057	1.009	−0.254***	0.092	0.696	−0.099	0.083	0.559	−0.335***	0.101	1.258
Treatment	−0.122**	0.054	0.333	−0.179***	0.033	0.912	−0.136**	0.056	0.373	−0.131***	0.048	0.743	−0.218***	0.059	0.818
Accident	−0.042	0.055	0.115	−0.082**	0.034	0.421	−0.046	0.058	0.126	0.022	0.049	-0.123	−0.230***	0.062	0.863
Noise pollution	0.082	0.062	-0.226	−0.135***	0.039	0.691	0.058	0.065	-0.158	−0.121**	0.055	0.687	−0.099	0.072	0.373
Air pollution	0.190***	0.061	-0.520	−0.087**	0.037	0.445	0.212***	0.065	-0.583	−0.042	0.053	0.240	−0.150**	0.068	0.561
Water pollution	−0.102	0.099	0.278	−0.033	0.052	0.168	0.013	0.105	-0.035	−0.048	0.072	0.270	−0.130	0.093	0.487
Garbage pollution	−0.038	0.065	0.104	0.016	0.040	-0.079	−0.035	0.068	0.095	0.099*	0.058	-0.559	−0.104	0.076	0.391
Robbery	−0.269***	0.051	0.739	0.039	0.031	-0.200	−0.278***	0.054	0.762	0.015	0.044	-0.087	0.071	0.056	-0.265
Poor surveillance	−0.110**	0.049	0.302	−0.032	0.031	0.164	−0.129**	0.051	0.354	−0.015	0.044	0.084	−0.051	0.055	0.190
Street drugs	−0.132**	0.054	0.363	−0.013	0.034	0.068	−0.157***	0.056	0.431	−0.030	0.049	0.167	0.014	0.062	-0.053
Drugs trafficking	0.243***	0.070	-0.667	−0.208***	0.041	1.064	0.262***	0.074	-0.718	−0.107*	0.058	0.605	−0.340***	0.074	1.276
Log(Density)	−0.049***	0.011	0.135	0.007	0.007	-0.034	−0.048***	0.012	0.131	−0.013	0.009	0.072	0.025*	0.013	-0.094
Log(Income)	0.251**	0.112	-0.689	−0.115*	0.059	0.586	0.270**	0.119	-0.742	−0.125*	0.075	0.706	−0.133	0.136	0.500
Urban	−0.002	0.051	0.004	0.122***	0.032	-0.621	−0.041	0.053	0.112	0.264***	0.047	-1.493	−0.037	0.057	0.138
x				4.622***	0.314					7.620***	0.380		5.135***	0.389	
y				−0.079***	0.013					−0.271***	0.017		−0.225***	0.017	
LL	-10490						-10451								
# parameters	48						73								
N	16008						16008								
AIC	21075						21049								
BIC	21181						21210								

Note: Standard errors in parentheses. $AIC = -2\log L + 2K$, $BIC = -2\log L + \log(N)K$. $^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.001$

differences, being a female in communes belonging to the class 2 would require an increase in household income pc of 35% to offset gender differences in well-being. While CV for gender differences is not significant (equal to zero) for the first class. Being unemployed is detrimental in both classes. However, the CV is higher in the second class. For example, individuals living in communes in the first class require an increase of their household income pc of about 88% in order to be as well-off as someone who is not unemployed compare with the 80% of the first class. Household size has similar CV in both classes, whereas being married is more valued in the second class.

- Health conditions (H_i): In terms of health conditions, disability is the detrimental effect that requires the highest monetary compensation among the three health conditions in both classes (68% in class 1 vs. 100% in class 2). Having some kind of medical treatment is significant in both classes, however individuals in communes of class 2 are more sensitive requiring a higher compensation compare to class 1 (33% vs. 91%). Finally, having an accident is, in average, only detrimental for communes in the second class requiring an increase of household income pc of about 42% in order to keep the same wellbeing as someone without an accident.

- Neighborhood characteristics (Z_i): In terms of perceived pollution in the neighborhood, water and trash pollution are not important. Those results are in line with those for the continuous spatial heterogeneity assuming correlated parameters and lognormal and truncated normal distribution. The detrimental effect of noise pollution is negative and significant for the second class, with a monetary compensation of about 69% increased in household income. Air pollution shows a more complex result. According the coefficient for the first class, individuals in those communes like more noise pollution resulting in a positive CV. It can be noticed that dis-amenities in the neighborhood such as robbery, poor surveillance and street drug problems are only important for communes in the first class. For example, individuals living in communes belonging to the first class need an increase of about 73%, 30% and 37% in household income pc respectively.

- Variables at the commune level: Density reduces subjective well-being of individuals in class 1: an increase on 1% on density would require an increase of 0.14% of household income to keep the individuals at the same level of subjective well-being. Log of median income at the commune level (our proxy for development) presents different results for both classes. For the first class, the log of median income presents a positive correlation with individuals' subjective well-being, whereas in the second class the correlation is negative. In other words, individuals living in communes of the first class enjoy better development so that they should be compensated by an increase of about 68% of household income per capita if median income is reduced by 1%. Oppositely, inhabitants of commune belonging in class 2 dislike higher levels of development; therefore they should be compensated with an increase of 59% of household income if median income increases by 1%. Finally, urbanism matters only in the second class. The benefit of moving from rural to urban communes is worth around a relative increase of 0.6 in household income for the second class.

3.4.5 Spatial heterogeneity in CV

In this Section I analyze the distribution of the conditional mean of compensating variations across communes estimated using Equations (3.12) and (3.13). Since the number of estimated parameters is large, it is almost impossible to comment in detail each coefficient. Given this restriction and the purpose of this study, I discuss some general aspects of the results, and then I will focus in CVs for certain variables.[23]

Figures 3.2-3.6 show the distribution (boxplot) of the conditional mean of CVs for each group of variables estimated using continuous and discrete heterogeneity. In particular, I plotted the distribution for the following models:

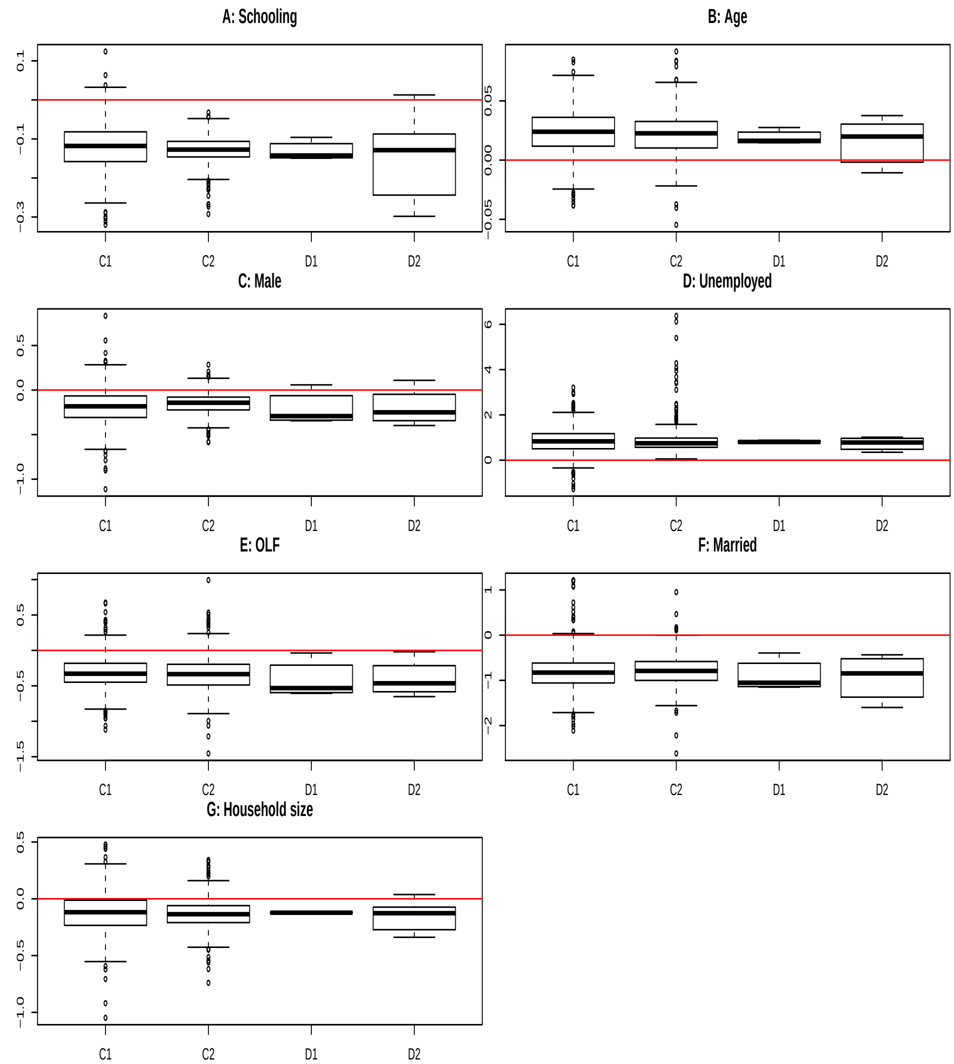
- C1: Continuous model with all parameters as normally distributed,
- C2: Continuous model with correlated random parameters (LN and CN),
- D1: Discrete model with $Q = 2$,
- D2: Discrete model with $Q = 3$.

The idea behind this exercise is to graphically detect differences in the CVs using both methods. The optimal scenario would be if we could observe similar distributional patterns of the compensating variations. This would indicate that the results are robust to any kind of spatial heterogeneity and assumption about the underlying distribution of the CVs. As we will see, this is not the case for this particular sample.

A first important result can be observed by taking a quick look to the distributions. When the parameters are assumed to be normally distributed, I always find communes with positive and negative compensating variations for all the variables. This result can be true or just an artifact of the normal distribution (recall that

[23]In cases where the research question is more direct, one could focus on particular communes if there are reasons to do so or describe the distribution of specific quantiles of the location-specific CVs as shown in Daziano and Achtnicht (2014).

Figure 3.2: Distribution of CVs for individual's covariates.



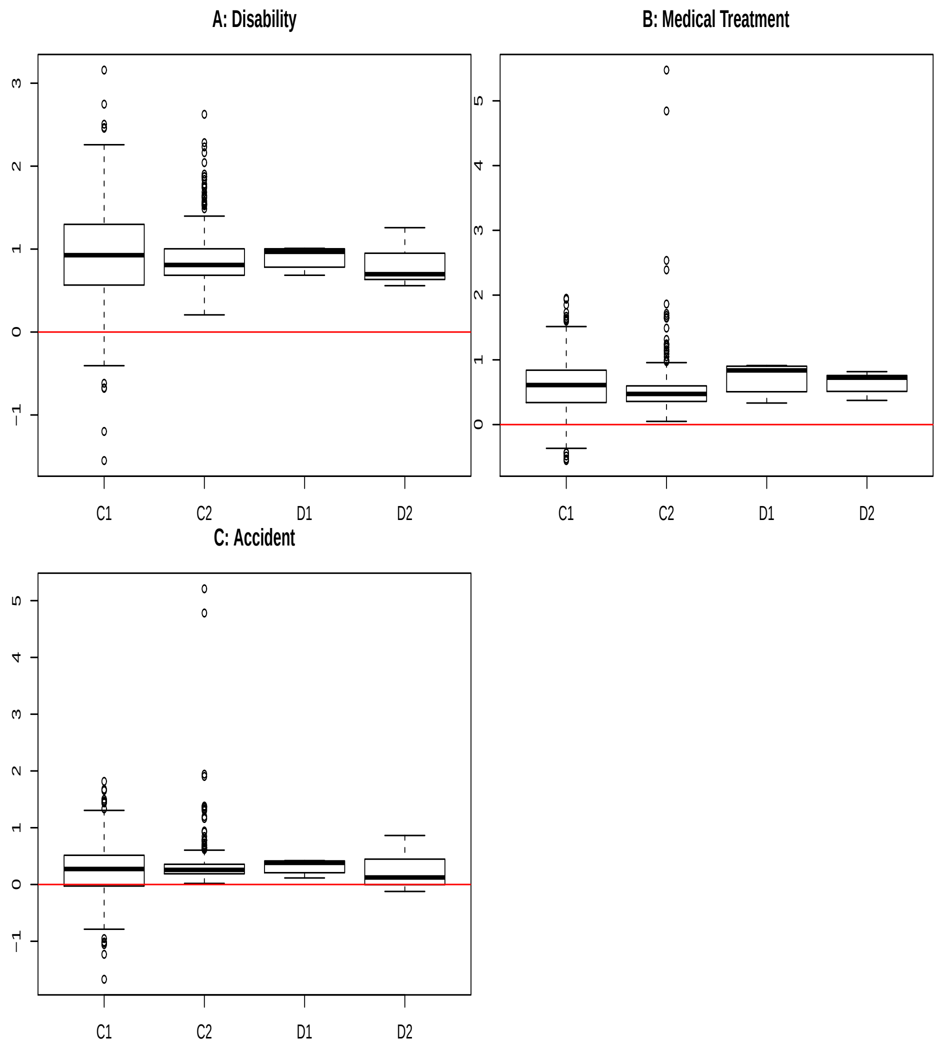
Notes: Each point correspond to the estimate of conditional expectation of compensating variation for each commune using Equation (3.12) and (3.13). The standard errors for the confidence intervals are computed using Equation (3.14). C1 correspond to CVs based on estimates using model 2 from Table 3.2; C2 correspond to CVs based on estimates using model 4 from Table 3.2; D1 and D2 correspond to CVs based on estimates using discrete distributions with 2 and 3 classes from Table 3.3.

the normal distribution assumes positive and negative coefficients whether this is true or not). For some of the variables, the latter seems to be true. For example, it is difficult to argue against the fact that disability has a detrimental effect on well-being, and therefore the compensatory variation in any geographical location must be positive. However, Panel A of Figure 3.3 shows that when the parameter is assumed to be normal distributed, there exists some communes with negative CV. The results are as (theoretically) expected when a discrete distribution is assumed or when it is assumed that the parameter is continuously distributed according to a log-normal distribution. A similar situation occurs when we look at treatment and accident.

In terms of variability, it can be observed that, in general, the discrete case shows lower variability. This is not unsurprising since the normal distribution (and any of its variations) has infinity tails, resulting in some communes having implausible extreme compensating variations, whereas the variability of the discrete case will depend on the number of classes.

Another important aspect is whether the location-specific CVs are significant. For illustration purposes, Figure 3.1 shows the compensating variations for disability for each commune along with the 95% confidence intervals (CI). The standard errors used to construct the CIs were computed using Equation (3.14). When looking at the continuous case, it can be observed that the compensating variations for some communes are not statistically significant since they contain zero. In fact, 78% and 72% of the communes have non-significant CV. Furthermore, when normally distributed parameters are considered, a small proportion of the communes exhibits unexpected negative CVs. These unexpected results are well-known problems of the normality assumption (see for example Daziano and Achtnicht, 2014). The log-normal assumption shows more credible results, since the CVs are forced to not contain negative values. For the discrete case we observed that in no case does any confidence interval include zero implying that CVs are statistically not zero and positive for all the communes (that is, individuals in all the communes

Figure 3.3: Distribution of CV for individual's health covariates.

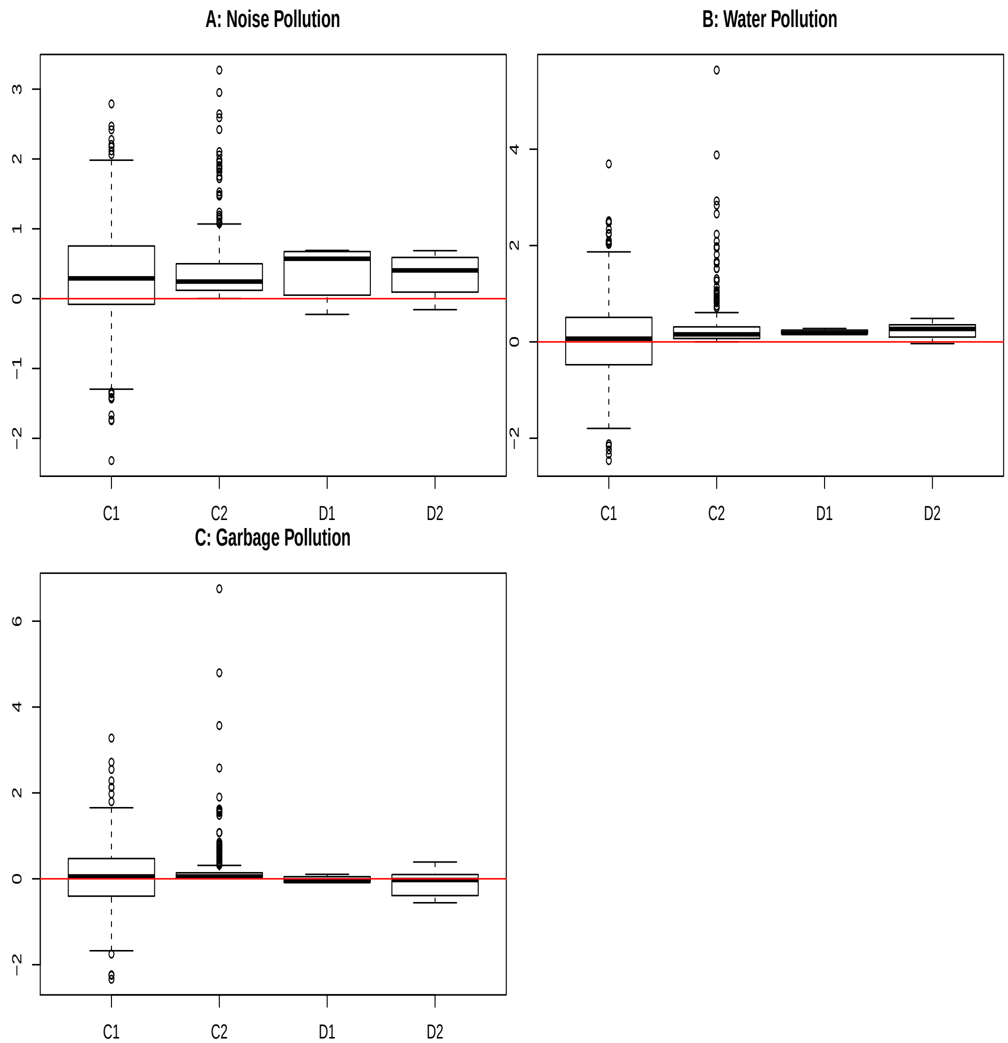


Notes: Each point correspond to the estimate of conditional expectation of compensating variation for each commune using Equation (3.12) and (3.13). The standard errors for the confidence intervals are computed using Equation (3.14). C1 correspond to CVs based on estimates using model 2 from Table 3.2; C2 correspond to CVs based on estimates using model 4 from Table 3.2; D1 and D2 correspond to CVs based on estimates using discrete distributions with 2 and 3 classes from Table 3.3.

should be compensated for the detrimental effect of disability).

Since non-pecuniary cost of unemployment on individual's subjective well-being is a very important topic on the SWB literature (Winkelmann and Winkelmann,

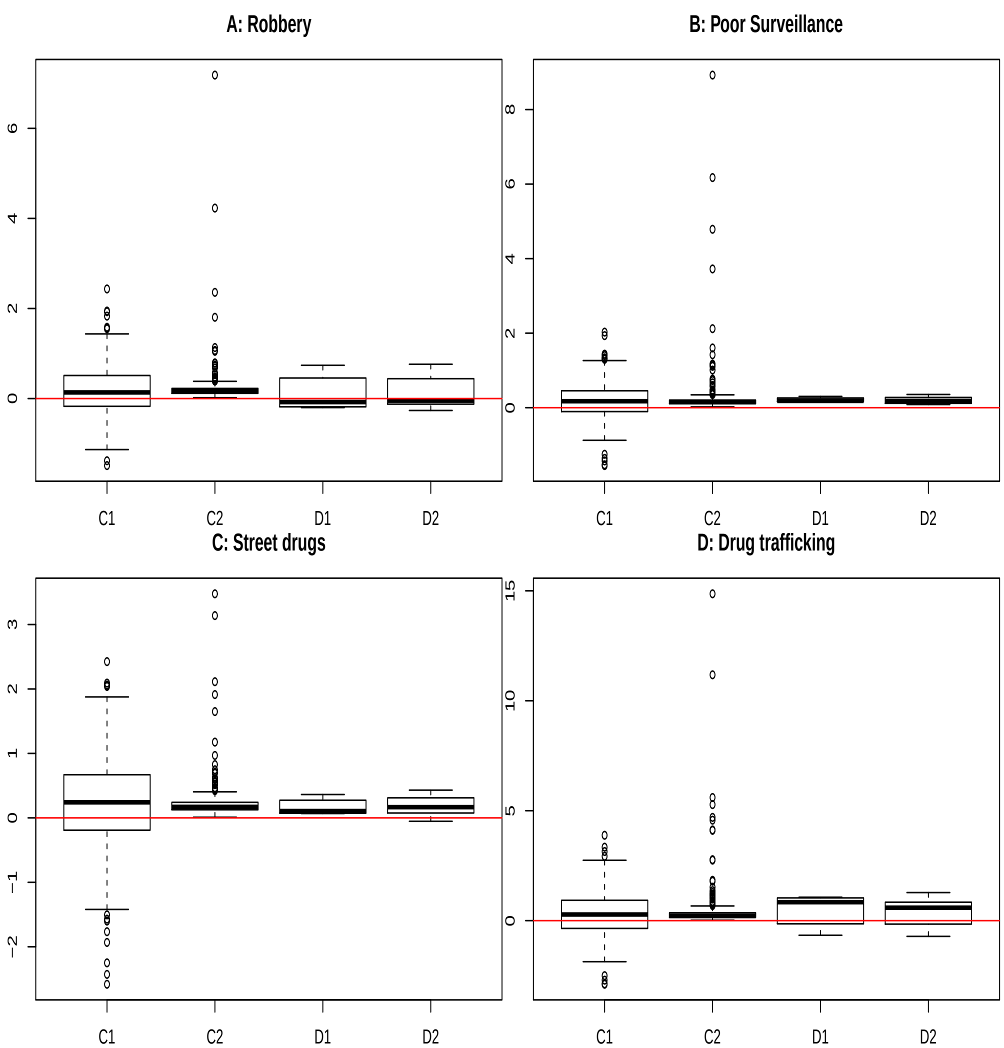
Figure 3.4: Distribution of CV for perceived neighborhood pollution.



Notes: Each point correspond to the estimate of conditional expectation of compensating variation for each commune using Equation (3.12) and (3.13). The standard errors for the confidence intervals are computed using Equation (3.14). C1 correspond to CVs based on estimates using model 2 from Table 3.2; C2 correspond to CVs based on estimates using model 4 from Table 3.2; D1 and D2 correspond to CVs based on estimates using discrete distributions with 2 and 3 classes from Table 3.3.

1998), Figure 3.8 plots confidence interval of compensating variation of unemployment for each commune. Again I find some communes with negative CV when assuming normal spatial heterogeneity, although those estimates are not significant. In effect, only 13% of the communes have significant CVs. When imposing

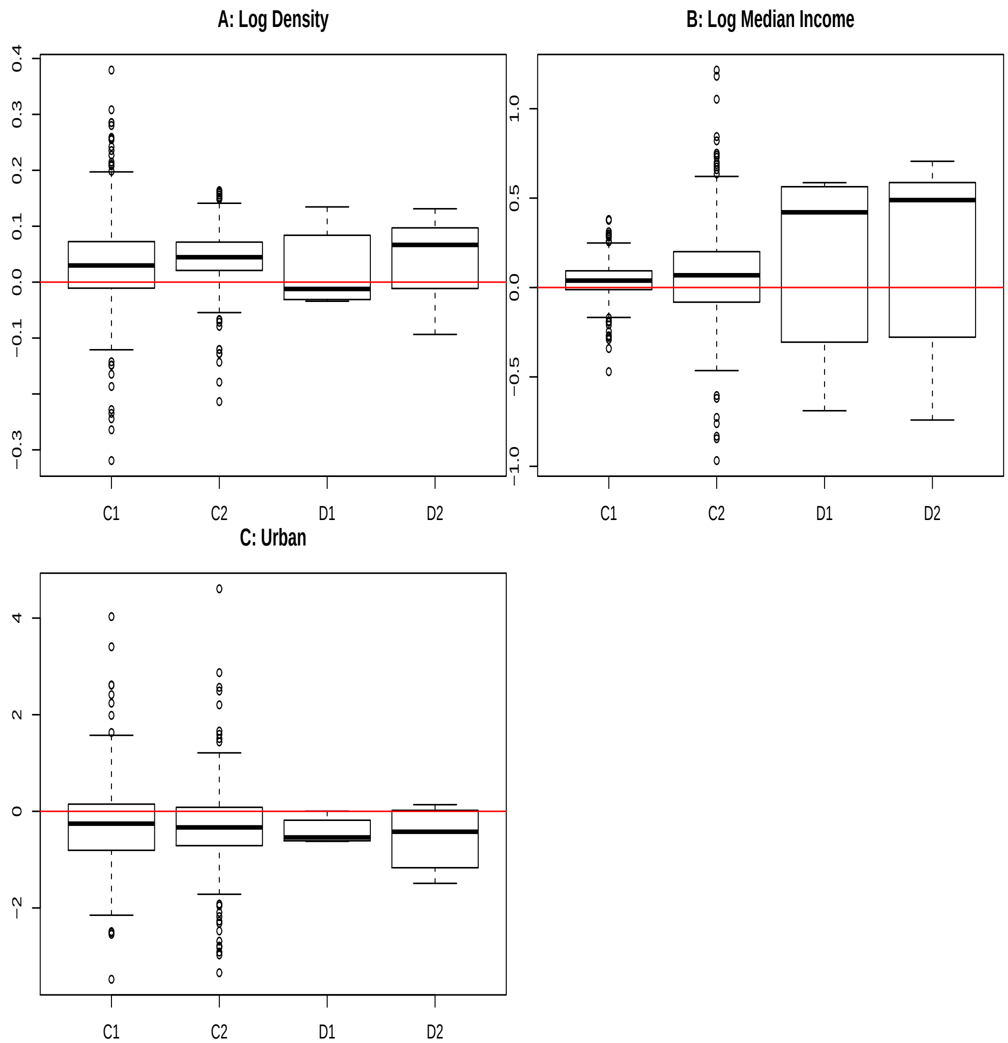
Figure 3.5: Distribution of CV for perceived public security in the neighborhood.



Notes: Each point correspond to the estimate of conditional expectation of compensating variation for each commune using Equation (3.12) and (3.13). The standard errors for the confidence intervals are computed using Equation (3.14). C1 correspond to CVs based on estimates using model 2 from Table 3.2; C2 correspond to CVs based on estimates using model 4 from Table 3.2; D1 and D2 correspond to CVs based on estimates using discrete distributions with 2 and 3 classes from Table 3.3.

the log-normal distribution, the proportion is reduced to 1%. The discrete case shows a similar pattern as the previous example: when assuming two classes all the communes have a significant CV respect to unemployment; whereas with three classes the proportion of communes with significant CV is roughly 82%.

Figure 3.6: Distribution of CV for variables at the commune-level.



Notes: Each point correspond to the estimate of conditional expectation of compensating variation for each commune using Equation (3.12) and (3.13). The standard errors for the confidence intervals are computed using Equation (3.14). C1 correspond to CVs based on estimates using model 2 from Table 3.2; C2 correspond to CVs based on estimates using model 4 from Table 3.2; D1 and D2 correspond to CVs based on estimates using discrete distributions with 2 and 3 classes from Table 3.3.

The lack of significance of location-specific CV under the continuous approach is striking, however it can be explained by considering the estimation procedure. The continuous approach, unlike the discrete approach, requires the simulations of the probability due to no close form solution. Therefore, in addition to the small-sample bias, the SML adds another layer of bias due to the simulation. In a Monte Carlo study, Sarrias (2015a) analyzes the ability of the conditional mean of the distribution of the spatially random parameters to retrieve the true spatial representation of the parameters under both methods. One of the main important results is that the continuous case requires more spatial units for the SML estimates to achieve acceptable levels of bias when compare with the ML estimates used in the discrete approach. In other words, given a small sample, the estimates from the discrete approach are more reliable than those from continuous approach.

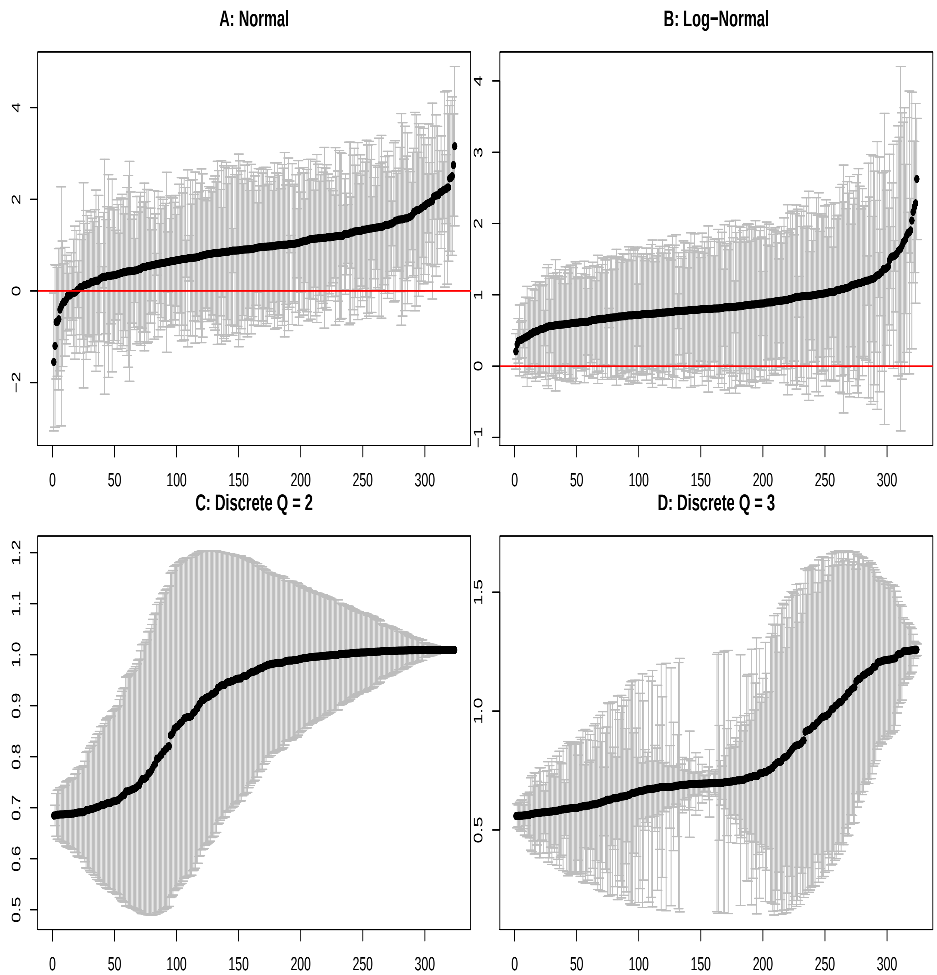
Another important issue is the number of parameters estimated under the more complex model. When assuming correlated parameters, the continuous model estimate 276 parameters compare with 73 parameters in the model with three classes. This might also have an impact on the variance of the estimated parameters.

3.5 Discussion and conclusion

Using SWB equations to estimate compensating variations has become a very simple but powerful tool for policy makers, especially for evaluating intangible goods that have no market price. Many of the theoretical and applied studies have helped us to expand our understanding of how people value different health conditions, environmental externalities, and their social capital. This study goes a step further and tries to add a new dimension to the debate by calling attention to the importance of spatial heterogeneity when analyzing compensating variations.

To this end, I compare two different models, continuous and discrete, to incorporate unobserved spatial heterogeneity into the estimation of CV. Both approaches assume that the coefficients (and therefore compensatory changes) follow

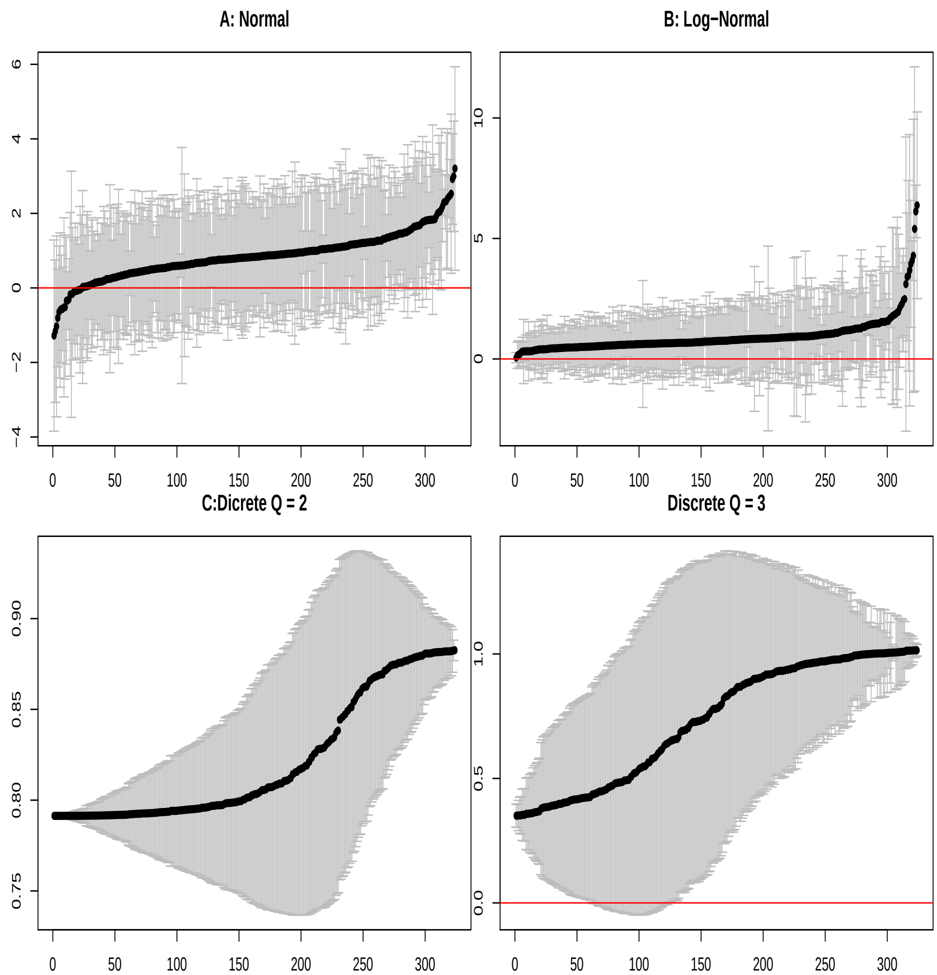
Figure 3.7: Compensating variation for disability with 95% CI.



Notes: Each point in the graph corresponds to a commune. The confidence intervals are computed using Equation (3.14). The graph from panel A corresponds to CVs estimated using coefficients from model 2 (All Normals) in Table 3.2. Panel B corresponds to CVs estimated using coefficients from model 3 (Correlated with LN and CN distributions) in Table 3.2 . Panel C and D correspond to CVs estimated using coefficient from models assuming discrete spatial heterogeneity. In particular, Panel C corresponds to model 1 in Table 3.3 and Panel D corresponds to model 2 in Table 3.3.

a distribution which is unknown to the investigator. In the first case, the spatial unobserved heterogeneity of coefficients is determined by a continuous distribution; therefore parameters may take any value within the domain of the chosen distribution. In the second case, spatial units (communes in this study) belong to different

Figure 3.8: Compensating variation for Unemployed with 95% CI.



Notes: Each point in the graph corresponds to a commune. The confidence intervals are computed using Equation (3.14). The graph from panel A corresponds to CVs estimated using coefficients from model 2 (All Normals) in Table 3.2. Panel B corresponds to CVs estimated using coefficients from model 3 (Correlated with LN and CN distributions) in Table 3.2 . Panel C and D correspond to CVs estimated using coefficient from models assuming discrete spatial heterogeneity. In particular, Panel C corresponds to model 1 in Table 3.3 and Panel D corresponds to model 2 in Table 3.3.

groups, and each group has the characteristic of having the same sensitivities to certain variables. Thus, there will be as many parameters for the same variable as classes of regions. In this study, the probability of belonging to a particular class is determined by the geographical location captured by the geographical coordinates.

Important findings emerged from this analysis applied to the Chilean case. The results support the evidence of substantial heterogeneity in the CV for several variables. For example, individuals perceiving poor surveillance in the neighborhood need in average a relative increase of household income pc of about 0.225 to be as well-off as someone who does not perceive this problem (figure based on the fixed model). However, when I allow for different compensating variations for each spatial unit I found that there are some communes where individuals should be compensated for much less, while for others communes more compensation is required. In other words, a similar negative characteristic of the neighborhood has varying compensating variation across communes. A non-stationary model would have disguised this remarkable geographical variation and led to broad generalizations ignoring these local differences in terms of CVs.

Although both approaches unveil spatial heterogeneity, I do find some differences on the results of both approaches. In this study, the discrete approach has a slight advantage over the continuous approach in terms of model fit: models with 2 and 3 classes present a better statistical fit compared to the non-stationary model and all the models with continuous spatial heterogeneity. This result is also found in similar studies that compare both methods in the context of Mixed and Latent Class Logit models (See for example Greene and Hensher, 2013; Shen, 2009; Hess et al., 2011). Furthermore, although the medians for the CVs with both approaches are relatively close, I noticed more mass at the extremes of the CV distributions when continuous spatial heterogeneity is assumed. This is partially explained by the domain of the continuous distributions assumed for the parameters. The normal distribution has infinite tails, which results in communes having implausible and extreme coefficients. Similarly, the main drawback of the log-normal distribution is that it has a very long right-hand tail. This characteristic of the log-normal distribution is corroborated for the health variables and neighborhood problems in Figures 3.3 and 3.5.

An important aspect of spatial analysis is whether CVs are significant for each location. Nevertheless, analyzing the evidence on CIs is challenging because of remarkably differences between both models. In general, more communes with significant CVs are found assuming a discrete rather than a continuous distribution. I speculate that this is due to the complexity of SML approach. In fact, Sarrias (2015a) performs a Monte Carlo experiment in order to understand the ability of both approaches to retrieve the true representation of the spatially varying process under small sample size. He finds that the data requirement to achieve lower bias in the continuous case is substantial compared with the discrete case. He also finds that the precision to identify each locational-specific estimate improves as the number of individuals per region increases. However, the discrete case is able to retrieve the true spatial heterogeneity surface with lower bias and better coverage for small samples.

So, the key question is which one should be used? Overall, the results of this study suggest that, at least for the sample used, the discrete approach provides better statistical fit and easier interpretation. In addition, the appeal of the discrete approach is that the analyst does not need to specify a particular distribution for the unobserved spatial heterogeneity. As I reviewed, assuming some continuous distribution might lead to unreasonable coefficients. In other words, when the distributions of the parameters are less constrained may lead to significantly better results. Furthermore, the SML approach used in the continuous case is very costly in terms of computational times. For example, the more complex model (model 4 of Table 3.2) took 1 hour and 29 minutes to converge, whereas the model with three classes took about 15 minutes. Nonetheless, this evidence is not conclusive and it can be sample-specific. Therefore, the need for ongoing research on the comparison between these two specifications with different samples is fundamental.

Finally, this work is not without shortcomings. First, the SWB approach needs very strong assumptions to be valid to proxy for individuals' utility. However, the literature reviewed in Section 3.2.1 suggests that using satisfaction with life (instead of other measures of SWB) is a more valid and reliable measure for measuring true well-being. Secondly, the results should be interpreted cautiously because of the potential endogeneity of some variables, especially income: while income may make people more satisfied with their life, inherently more satisfied people may earn higher income (Levinson, 2012). Powdthavee (2010a) and Luttmer (2005) use instrumental variable approach and both find that income coefficient in IV specifications are larger than in models that assume exogeneity. These results, along with the Benjamin et al. (2012)'s results under an experimental setting, suggest that CVs given by Equation (3.4) are overstated. Given this, I confine myself to talk about spatial heterogeneous relationships instead of causation acknowledging this potential upward bias. Nevertheless, one can think of the study as adopting strong identification assumption, but exploring how we can push their findings further by exploring spatial heterogeneity.

Despite these limitations, this study has important implications. I argue that when analyzing CVs explicit recognition of spatial heterogeneity is required. I have shown that the response of individuals' underlying utility to covariates is not fixed over space, but rather it varies among space in terms of intensity. From a policy perspective, an understanding of how certain geographical areas need more ore less compensation can be helpful in designing effective geographical interventions.

BIBLIOGRAPHY

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.

Anastasopoulos, P. C. and Mannering, F. L. (2009). A Note on Modeling Vehicle Accident Frequencies with Random-Parameters Count Models. *Accident Analysis & Prevention*, 41(1):153–159.

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.

Benjamin, D., Heffetz, O., Kimball, M., and Rees-Jones, A. (2014). Can Marginal Rates of Substitution Be Inferred From Happiness Data? Evidence from Residency Choices. *The American economic review*, 104(11):3498–3528.

Benjamin, D. J., Kimball, M. S., Heffetz, O., and Rees-Jones, A. (2012). What Do You Think Would Make You Happier? What Do You Think You Would Choose? *The American economic review*, 102(5):2083.

Bhat, C. R. (2001). Quasi-random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model. *Transportation Research Part B: Methodological*, 35(7):677–693.

Bockstael, N. E. and Freeman, A. M. (2005). Welfare Theory and Valuation. In Mler, K.-G. and Vincent, J. R., editors, *Valuing Environmental Changes*, volume 2 of *Handbook of Environmental Economics*, pages 517 – 570. Elsevier.

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.

Bruni, L. (2007). *Handbook on the Economics of Happiness*. Edward Elgar Publishing.

Bruni, L. and Porta, P. L. (2005). *Economics and Happiness: Framing the Analysis*. Oxford University Press.

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.

Clark, A. E. and Oswald, A. J. (2002). A Simple Statistical Method for Measuring How Life Events Affect Happiness. *International Journal of Epidemiology*, 31(6):1139–1144.

Croissant, Y. (2013). **pglm***: Panel Generalized Linear Model*. R package version 0.1-2.

Croissant, Y. et al. (2012). Estimation of Multinomial Logit Models in R: The **mlogit** Packages. *R package version 0.2-2*.

Daly, A., Hess, S., and Train, K. (2012). Assuring Finite Moments for Willingness to Pay in Random Coefficient Models. *Transportation*, 39(1):19–31.

Daziano, R. A. and Achtnicht, M. (2014). Accounting for Uncertainty in Willingness to Pay for Environmental Benefits. *Energy Economics*, 44:166–177.

Deiner, E., Suh, E., Lucas, R. E., and Smith, H. L. (1999). Subjective Well-being: Three Decades of Progress. *Psychological Bulletin*, 125(2):276–302.

Di Tella, R. and MacCulloch, R. (2006). Some Uses of Happiness Data in Economics. *The Journal of Economic Perspectives*, pages 25–46.

Diener, E., Suh, E. M., Smith, H., and Shao, L. (1995). National Differences in Reported Subjective Well-being: Why Do They Occur? *Social Indicators Research*, 34(1):7–32.

Doran, H., Bates, D., Bliese, P., Dowling, M., et al. (2007). Estimating the Multilevel Rasch Model: With the **lme4** Package. *Journal of Statistical Software*, 20(2):1–18.

Dumont, J., Keller, J., and Carpenter, C. (2014). **RSGHB***: Functions for Hierarchical Bayesian Estimation: A Flexible Approach*. R package version 1.0.2.

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.

Eid, M. and Larsen, R. J. (2008). *The Science of Subjective Well-Being*. Guilford Press.

Elff, M. (2014). **memisc***: Tools for Management of Survey Data, Graphics, Programming, Statistics, and Simulation*. R package version 0.96-10.

Falco, P., Maloney, W. F., Rijkers, B., and Sarrias, M. (2015). Heterogeneity in Subjective Wellbeing: An Application to Occupational Allocation in Africa. *Journal of Economic Behavior & Organization*, 111(0):137 – 153.

Ferrer-i Carbonell, A. and Frijters, P. (2004). How Important is Methodology for the Estimates of the Determinants of Happiness? *The Economic Journal*, 114(497):641–659.

Ferrer-i Carbonell, A. and van Praag, B. (2002). The Subjective Costs of Health Losses due to Chronic Diseases. An Alternative Model for Monetary Appraisal. *Health economics*, 11(8):709.

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.

Fox, J., Bates, D., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Monette, G., Nilsson, H., Ogle, D., et al. (2009). **CAR***: Companion to Applied Regression, R Package Version 1.2-16*.

Frey, B. S. (2008). Happiness: A Revolution in Economics. *MIT Press Books*, 1.

Frey, B. S., Luechinger, S., and Stutzer, A. (2010). The Life Satisfaction Approach to Environmental Valuation. *The Annual Review of Resource Economics*, 2:139–60.

Frey, B. S. and Stutzer, A. (2002). What Can Economists Learn from Happiness Research? *Journal of Economic literature*, pages 402–435.

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.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo Maximum Likelihood Methods: Applications to Poisson Models. *Econometrica: Journal of the Econometric Society*, 52(3):701–720.

Greene, W. (2007). *Functional Form and Heterogeneity in Models for Count Data*. Now Publishers Inc.

Greene, W., Harris, M. N., and Spencer, C. (2014). Estimating the Standard Errors of Individual-Specific Parameters in Random Parameters Models. Technical report.

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

Greene, W. H. (2015a). *LIMDEP: Version 10: Econometric Modeling Guide*. Econometric Software.

Greene, W. H. (2015b). *NLOGIT: Version 5: User's Guide*. Econometric Software.

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.

Greene, W. H. and Hensher, D. A. (2010a). *Modeling Ordered Choices: A Primer.* Cambridge University Press.

Greene, W. H. and Hensher, D. A. (2010b). Ordered Choices and Heterogeneity in Attribute Processing. *Journal of Transport Economics and Policy (JTEP)*, 44(3):331–364.

Greene, W. H. and Hensher, D. A. (2013). Revealing Additional Dimensions of Preference Heterogeneity in a Latent Class Mixed Multinomial Logit Model. *Applied Economics*, 45(14):1897–1902.

Groot, W. and Maassen van den Brink, H. (2006). The Compensating Income Variation of Cardiovascular Disease. *Health Economics*, 15(10):1143–1148.

Grün, B. and Leisch, F. (2008). **FlexMix** Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *Journal of Statistical Software*, 28(4):1–35.

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

Hajivassiliou, V. A. and Ruud, P. A. (1994). Classical Estimation Methods for LDV Models Using Simulation. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume 4 of *Handbook of Econometrics*, chapter 40, pages 2383–2441. Elsevier.

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

Henningsen, A. (2014). **micEcon***: Microeconomic Analysis and Modelling.* R package version 0.6-12.

Henningsen, A. and Toomet, O. (2011). **maxLik**: A Package for Maximum Likelihood Estimation in R. *Computational Statistics*, 26(3):443–458.

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

Hess, S. (2014). Latent Class Structures: Taste Heterogeneity and Beyond. *Handbook of Choice Modelling*, page 311.

Hess, S., Ben-Akiva, M., Gopinath, D., and Walker, J. (2011). Advantages of Latent Class Over Continuous Mixture of Logit Models. *Institute for Transport Studies, University of Leeds. Working paper.*

Jackson, C. H. (2011). Multi-state Models for Panel Data: The **msm** Package for R. *Journal of Statistical Software*, 38(8):1–29.

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.

Krapf, M., Ursprung, H. W., and Zimmermann, C. (2014). Parenthood and Productivity of Highly Skilled Labor: Evidence From the Groves of Academe. Technical report, FRB of St. Louis Working Paper.

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.

Levinson, A. (2012). Valuing Public Goods Using Happiness Data: The Case of Air Quality. *Journal of Public Economics*, 96(9):869–880.

Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M., and Richardson, S. (2015). **PReMiuM**: An R Package for Profile Regression Mixture Models Using Dirichlet Processes. *Journal of Statistical Software*, 64(7):1–30.

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

Long, J. S. (1990). The Origins of Sex Differences in Science. *Social Forces*, 68(4):1297–1316.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*, volume 7. Sage.

Luttmer, E. F. (2005). Neighbors as Negatives: Relative Earnings and Well-Being. *The Quarterly journal of economics*, 120(3):963–1002.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York.

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

Mentzakis, E. (2011). Allowing for Heterogeneity in Monetary Subjective Well-being Valuations. *Health economics*, 20(3):331–347.

Moore, S. and Shepherd, J. P. (2006). The Cost of Fear: Shadow Pricing the Intangible Costs of Crime. *Applied Economics*, 38(3):293–300.

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.

Perez-Truglia, R. (2015). A Samuelsonian Validation Test for Happiness Data. *Journal of Economic Psychology*, 49:74–83.

Powdthavee, N. (2008). Putting a Price Tag on Friends, Relatives, and Neighbours: Using Surveys of Life Satisfaction to Value Social Relationships. *The Journal of Socio-Economics*, 37(4):1459–1480.

Powdthavee, N. (2010a). How Much Does Money Really Matter? Estimating the Causal Effects of Income on Happiness. *Empirical Economics*, 39(1):77–92.

Powdthavee, N. (2010b). *The Happiness Equation: The Surprising Economics of our Most Valuable Asset*. Icon Books.

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.

Sándor, Z. and Train, K. (2004). Quasi-random Simulation of Discrete Choice Models. *Transportation Research Part B: Methodological*, 38(4):313–327.

Sarrias, M. (2015a). Continuous and Discrete Spatial Heterogeneity: Modeling Strategies and Simulation. Technical report.

Sarrias, M. (2015b). **Rchoice***: Discrete Choice (Binary, Poisson and Ordered) Models with Random Parameters*. R package version 0.3.

Sarrias, M. and Daziano, R. (2015). **gmnl***: Multinomial Logit Models with Random Parameters*. R package version 1.0.

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.

Shen, J. (2009). Latent Class Model or Mixed Logit Model? A Comparison by Transport Mode Choice Data. *Applied Economics*, 41(22):2915–2924.

StataCorp (2011). STATA Statistical Software: Release 12. *College Station, TX: StataCorp LP*.

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. (2000). Halton Sequences for Mixed Logit. *Department of Economics, UCB*.

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

Van Praag, B. and Baarsma, B. E. (2005). Using Happiness Surveys to Value Intangibles: The Case of Airport Noise. *The Economic Journal*, 115(500):224–246.

Van Praag, B. M. (2007). Perspectives from The Happiness Literature and the Role of New Instruments for Policy Analysis. *CESifo Economic Studies*, 53(1):42–68.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer-Verlag, New York, 4th edition.

Welsch, H. and Kühling, J. (2009). Using Happiness Data for Environmental Valuation: Issues and Applications. *Journal of Economic Surveys*, 23(2):385–406.

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.

Winkelmann, L. and Winkelmann, R. (1998). Why Are The Unemployed so Unhappy? Evidence from Panel Data. *Economica*, 65(257):1–15.

Winkelmann, R. and Boes, S. (2006). *Analysis of Microdata*. Springer.

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

Zeileis, A. (2006). Object-Oriented Computation of Sandwich Estimators. *Journal of Statistical Software*, 16(i09).

Zeileis, A. and Hothorn, T. (2002). Diagnostic Checking in Regression Relationships. *R News*, 2(3):7–10.