2013•2014
FACULTY OF SCIENCES
Master of Statistics

Master's thesis

Use of Bayesian variable selection methods in spatial regression models

Promotor :
Prof.dr. Christel FAES

Tapiwa Ganyani

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

universiteit hasselt
KNOWLEDGE IN ACTION

Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

universiteit hasselt | Maastricht University

2013•2014
FACULTY OF SCIENCES
Master of Statistics

Master's thesis

Use of Bayesian variable selection methods in spatial regression models

Promotor :
Prof.dr. Christel FAES

Tapiwa Ganyani
Thesis presented in fulfillment of the requirements for the degree of Master of Statistics

universiteit hasselt | Maastricht University

Acknowledgments

First of all, I would like to thank God the Almighty for granting me the opportunity to undertake master studies. I also thank him for giving me strength and guidance to accomplish this thesis project.

I am very grateful to Vlaamse Interuniversitaire Raad (VLIR-UOS) for the funding, your financial support made it possible for me to fully focus on my studies. I will always feel honored to be a recipient of the VLIR scholarship.

I would like to express my sincere appreciation to my thesis supervisor Prof. dr. Christel Faes for giving support and insightful comments from the start to the end of this thesis project. I would also like to extend my appreciation to all Professors who taught me during my master studies, the knowledge and skills you gave me helped a lot in this thesis project.

Abstract

The Thyroid Stimulating Hormone (TSH) is a hormone that plays a role in brain development and bone maintenance among other functions. Lack of dietary iodine is known to result in decreased production of thyroid hormones and the deficiency of these hormones triggers a rise in blood TSH levels. Though iodine influences TSH levels, there could be other important predictor variables as well. In this project we consider four potential predictors of TSH levels in newborns (gestational age, birth weight, gender and time until first feeding) within a spatial regression context and the idea is to select which ones are important to include in the regression model. We illustrate and compare some Bayesian variable selection methods using two types of spatial regression models namely, spatially fixed coefficient models and spatially varying coefficient models.

Contents

List of Figures

List of Tables

1 Introduction

The Thyroid Stimulating Hormone (TSH) is a hormone whose function is to regulate thyroid hormones which are important in brain development and bone maintenance among other functions. Lack of dietary iodine leads to reduced production of thyroid hormones and a deficiency of these hormones triggers a rise in blood TSH levels. The increased levels of TSH are aimed at stimulating the thyroid to produce more thyroid hormones into the blood so as to return thyroid hormone levels to normal state. Upon stimulation, many biochemical reactions occur in the thyroid and eventually its cells grow and multiply leading to goitre (American Thyroid Association, 2012).

Iodine deficiency is more common in mountainous regions of the world where food is grown in iodine-poor soil. In these regions low content of soil iodine is due to leaching effects of snow, water and heavy rainfall, which removes iodine from the soil (Assey *et al*, 2006). If iodine does not enter the dietary chain of populations living in these regions (via e.g. iodization of salt) then iodine deficiency disorders will persist in these areas (World Health Organization, 2007).

Even though iodine influences TSH levels, there could be other predictor variables as well. In this project we consider some potential predictors of TSH within a regression setting and the idea is to select which ones to include in the regression model. This exercise is known as variable selection and can be seen as a way to identify important and negligible predictors (Mitchell and Beauchamp, 1998).

We base our study on a community called Galicia (Spain) which consists of several municipalities some of which are severely affected by iodine deficiency (Lope *et al*, 2006). It is worth noting that there may be similarity in TSH levels amongst municipalities which are closer together as compared to municipalities that are further apart possibly due to the amount of iodine content in the soil, dietary habits as well as other factors. Hence, we consider spatial regression models.

In spatial regression modelling a common approach is to assume that effects of predictor variables are the same for all areas, and thus a single effect for each predictor variable is estimated for the entire region. In order to select predictor variables to include in the regression model, one may utilize model selection tools to compare across models consisting of a different subset of predictor variables. In a Bayesian context (the context in which spatial modelling is often done), one may use the Deviance Information Criterion (DIC), Bayes Factor and Mean Square Prediction Error among others. However, a drawback of using these tools is that they are calculated for every model to be compared and if there are many predictor variables to be considered the exercise becomes laborious. In view of this problem, Bayesian Variable Selection (BVS) techniques come in handy - they save one from the laborious exercise via automating the variable selection process.

Instead of assuming that effects of predictor variables are fixed over the entire region, one can assume that these effects vary from one area to another. Here, the model to be fitted will contain regression coefficients for each predictor variable at each area, in other words a regression model for each area is fitted. As is the case with spatially fixed covariate effects, the problem of selecting predictor variables which are important in explaining the responses arises. It may be that a given predictor variable is important at one area and not important at another. Thus we can again use BVS techniques with the area specific regression models to identify which predictor variables are important for a given area.

In doing the area-specific variable selection it may be reasonable to assume that if a predictor variable is important at a given area, then it would also be important among its neighbours. Hence, one may consider incorporating spatial structure in the variable selection process (Lum, 2012).

In this project, we illustrate and compare BVS techniques in the context of spatially fixed and spatially varying coefficient models. In particular, we apply the method of Kuo and

Mallick (1998), Gibbs Variable Section (Dellaportas *et al*, 1997) and Stochastic Search Variable Selection (George and McCulloch, 1993).

1.1 Study objectives

The objective of this project is to investigate how and whether BVS techniques can be utilized in spatial regression models.

1.2 Description of the Data

Galicia is located in the northwest of Spain and has a total area of 29, 574km^2. Its interior is characterized by a mountainous landscape and in terms of climate it is temperate and rainy. Galicia therefore suffers from iodine deficiency.

The data considered in this project are TSH levels of 15,416 babies born in 315 municipalities of Galicia (Spain) in the year 2009. TSH screening was done on all these newborns, approximately 3 days after birth. During screening, some variables were recorded together with level of TSH namely, gestation age (in weeks), birth weight (in grams), gender and time from birth until first feeding (in days).

2 Methodology

2.1 Model Specification

Let $Y_i^{(j)}$ be a continuous response and $(X_{1i}^{(j)}, X_{2i}^{(j)}, \ldots, X_{pi}^{(j)})$ be a set of potential predictor variables for the i^{th} subject in area j $(i = 1, 2, \ldots, n; j = 1, 2, \ldots, m)$ and p denotes the number of predictor variables. Interest lies on modelling $Y_i^{(j)}$ as a linear function of the $X_i^{(j)}s$ as follows:

$$\begin{aligned} Y_i^{(j)} &= \beta_0 + \beta_1 X_{1i}^{(j)} + \cdots + \beta_p X_{pi}^{(j)} + \varepsilon_i^{(j)} + u^{(j)} + v^{(j)} \\ &= \beta_0 + \sum_{k=1}^{p} \beta_k X_{ki}^{(j)} + \varepsilon_i^{(j)} + u^{(j)} + v^{(j)} \end{aligned} \tag{1}$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are fixed regression coefficients; $\varepsilon_i^{(j)} \sim N(0, \sigma^2)$ is the error term for subject i in area j; σ^2 is the variance of the $\varepsilon_i^{(j)}$ terms; $u^{(j)} \sim N(0, \sigma_u^2)$ is the non-spatial random effect for area j; σ_u^2 is the variance of the $u^{(j)}s$; $v^{(j)}|v^{(m)}, m \in \delta_j \sim N(\bar{v}^{(j)}, \sigma_v^2)$ is the area j spatial random effect; δ_j is the set of neighbours of area j, $\bar{v}^{(j)}$ denotes the average of the v terms of the neighbours of area j; σ_v^2 is the variance of the $v^{(j)}|v^{(m)}s$.

According to Besag *et al* (1991), the terms u and v, are interpreted as surrogates for unknown or unobserved predictor variables at the area level; the $u's$ represent unstructured variables whereas the $v's$ represent structured variables. If the $v's$ were to be observed, they would display substantial spatial structure in that the values of neighbours would be generally similar. Hence, the $v's$ are assigned a Conditional Autoregressive (CAR) prior as one would think that the full conditional distribution of $v^{(j)}$ should depend only on its neighbours (Banerjee *et al*, 2004).

In this project we consider a fully Bayesian implementation of this model in WinBUGS. In terms of prior distributions we assign minimally informative priors on $\beta_1, \beta_2, \ldots, \beta_p$, i.e. $N(0, \tau_k^2)$, with τ_k^2 very large and IG(ϵ, ϵ) with ϵ small. Note that an improper uniform

prior is assigned on β_0 for the reason that the CAR distribution (*car.normal*) in WinBUGS is parameterised to include a sum-to-zero constraint on the spatial random effects.

In studies where the main goal is to obtain a good understanding between $Y_i^{(j)}$ and the $X_i^{(j)}s$, variable selection is a key step. It is key because it facilitates selection of predictor variables which are essential in explaining the response variable - this way, a better understanding of the Y/X relationship is obtained.

In subsection 2.2 we reflect on classical variable selection and in subsection 2.3 we give a detailed description of some Bayesian Variable Selection (BVS) techniques. In subsection 2.4 we describe BVS as extended to spatially varying coefficient models.

2.2 Classical Variable Selection

We use the term classical variable selection to refer to Bayesian model selection tools which overlap in use to the variable selection tool kit. Model selection encompasses identifying a 'good' model among competing models where these models differ in any of the following: link function, systematic component and/or random component. Since variable selection entails identifying a 'good' model through varying the systematic component to include different subsets of predictor variables it can be viewed as a special case of model selection. In this subsection we reflect on the DIC because our main focus is on BVS and also because of its popularity as a model selection tool.

2.2.1 Deviance Information Criterion

The Deviance Information Criterion (DIC) can be seen as a Bayesian version of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) in the sense that it trades off a measure of model adequacy against a measure of complexity. It is defined as

$$DIC = D(\bar{\boldsymbol{\beta}}) + 2p_D$$

where $D(.)$ is the deviance and $\bar{\boldsymbol{\beta}}$ is the posterior mean vector of unknown parameters of

the model. The deviance measures how well the model fits to the data, with larger values corresponding to a worse model. The parameter p_D is the effective number of parameters and measures the complexity of the model. This parameter is defined as the difference between posterior mean of the deviance and the deviance evaluated at the posterior mean of the parameters, i.e., $\overline{D(\boldsymbol{\beta})} - D(\bar{\boldsymbol{\beta}})$.

A model is penalized by both the value of $\overline{D(\boldsymbol{\beta})}$, which favors a good fit and by p_D. The quantity $\overline{D(\boldsymbol{\beta})}$ will decrease as the number of parameters in a model increases, p_D compensates for this effect by favoring models with a smaller number of parameters (Spiegelhalter *et al*, 2002). A rule of thumb for using DIC is that a difference in DIC of more than 10 rules out the model with the higher DIC while with a difference of less than 5 there is no clear winner (Lesaffre and Lawson, 2012).

Comparison of models via DIC involves calculating this measure for each model to be compared. A major drawback of such an exercise is that the number of possible models quickly grows large as the number of covariates increases - with p predictors, the number of possible models is 2^p. When p is large it becomes essential to employ variable selection strategies that can quickly identify promising models without having to fit all 2^p models.

2.3 Bayesian Variable Selection

In Bayesian literature there exist variable selection tools which possess the key feature of circumventing the computational burden which may be encountered with use of classical variable selection tools. Instead of calculating a variable selection measure for each of the 2^p models, these techniques make use of Monte Carlo Markov Chain (MCMC) sampling to quickly identify high probability models or in other words more promising models. Most of the unpromising submodels will never or rarely appear in the MCMC sample. As a precursor to the description of these techniques we give a brief account of the principle behind BVS.

Let $\mathcal{M}$ be the set of all 2^p models under consideration and m denote a model which is a member of this set. Also, let $f(m)$ denote the prior probability of model m and $f(\boldsymbol{y}|\boldsymbol{\beta}_m, m)$ the likelihood of the data under model m then, invoking the Bayes Theorem we have that the posterior probability of model m is given by:

$$f(m|\boldsymbol{y}) = \frac{f(m)f(\boldsymbol{y}|m)}{\sum_{m\in\mathcal{M}} f(m)f(\boldsymbol{y}|m)} \tag{2}$$

with

$$f(\boldsymbol{y}|m) = \int f(\boldsymbol{y}|\boldsymbol{\beta}_m, m)f(\boldsymbol{\beta}_m|m)d\boldsymbol{\beta}_m \tag{3}$$

where $\boldsymbol{\beta}_m$ denotes the regression coefficients of model m (as in model 1 above) (Dellaportas *et al*, 1997).

The principle behind BVS is to calculate the posterior model probabilities using expression (2) such that a model with high probability is chosen as the one containing essential predictor variables, this model is also known as the *maximum a posterior* (MAP) model. However Barbieri and Berger (2004) show that the MAP model is not necessarily the optimal predictive model and to this end, they show that the optimal predictive model is often the median probability (MP) model. The MP model is defined as the model where each predictor variable has posterior inclusion probability greater or equal to 0.5. In contrast to the MAP model which involves monitoring model posterior probabilities, the MP involves monitoring posterior inclusion probabilites for each covariate. In this project we applied and compared both the MAP and MP model approaches.

Several Bayesian variable selection methods have been proposed in the literature, these include among others: Kuo and Mallick, Gibbs Variable Selection (GVS), Stochastic Search Variable Selection (SSVS), adaptive shrinkage with Jeffreys' prior or a Laplacian prior, reversible jump MCMC and variable selection based on Zellner's g-prior (Kuo and Mallick, 1998; Dellaportas *et al*, 1997; George and McCulloch, 1997; O'Hara and Sillan-

paa, 2009; Lesaffre and Lawson, 2012). Here, three of these apporaches are considered and investigated in the context of spatial regression models, namely Kuo and Mallick, GVS and SSVS. These three are based on the Gibbs sampler and can therefore be easily implemented in WinBUGS software (Ntzoufras, 2009). Description of these methods follows.

2.3.1 Kuo and Mallick

Let $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)$ be a $p \times 1$ vector of latent indicators (also known as entry parameters), where the parameter γ_k is an indicator variable taking on values 1 or 0 when covariate X_k is included or excluded (respectively) in the model, with $k \in 1, 2, \ldots, p$. In this approach the linear predictor $\eta_i^{(j)} = \sum_{k=1}^{p} \beta_k X_{ki}^{(j)}$ of model (1) is expanded to include covariate entry parameters as follows:

$$\eta_i^{(j)} = \sum_{k=1}^{p} \gamma_k \beta_k X_{ki}^{(j)}. \tag{4}$$

Since estimation of the model is done in a Bayesian framework we specify a prior distribution on each of the parameters as was done for model (1). This method assumes that β and γ are independent a priori, i.e., $f(\beta_k, \gamma_k) = f(\beta_k) f(\gamma_k)$ and therefore implementation of this approach only requires specification of priors on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ independently (Kuo and Mallick, 1998). Prior distributions were specified as follows:

i. $\gamma_k \sim Bern(p_k)$ with $p_k \sim U(0, 1)$ for $k = 1, 2, \ldots, p$;

ii. $\beta_0 \sim dflat()$, $\beta_k \sim N(0, 10)$ for $k = 1, 2, \ldots, p$; and

iii. IG$(0.001, 0.001)$ for each of σ^2, σ_u^2 and σ_v^2.

The parameter p_k is the prior inclusion probability of the k^{th} predictor variable, i.e., the probability that predictor variable X_k is important a priori. The parameter is assigned a $U(0, 1)$ distribution to reflect no prior choice for the value p_k. Note that this specification assumes a mixture of a point mass at 0 with probability $1 - p_k$ and a normal density $N(.,.)$ with probability p_k for the parameter $\vartheta_k = \gamma_k \beta_k$. In this way, variable selection

is essentially a discrete process where covariates are either retained or dropped from the model. Note that we also chose to work with weakly informative priors on the $\beta_k s$ and the variance components as no prior information was available a priori.

The Gibbs sampler full conditionals for β_k are equal to:

$$f(\beta_k|\boldsymbol{y},\boldsymbol{\gamma},\boldsymbol{\beta}_{(k)}) \propto \begin{cases} f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\gamma})f(\beta_k|\boldsymbol{\beta}_{(k)}) & \text{if } \gamma_k = 1, \\ f(\beta_k|\boldsymbol{\beta}_{(k)}) & \text{if } \gamma_k = 0. \end{cases}$$

where $\boldsymbol{\beta}_{(k)}$ and $\boldsymbol{\gamma}_{(k)}$ are the vector of regression coefficients/entry parameters without the k^{th} component (Dellaportas *et al*, 1997). Clearly it can be seen from the full conditionals that when $\gamma_k = 0$, β_k will be sampled from the full conditional distribution $f(\beta_k|\boldsymbol{\beta}_{(k)})$, which is its prior distribution. On this note O'Hara and Sillanpaa (2009) point out that mixing will be poor if this prior is too vague as the sampled values of β_k will rarely be in the region where ϑ_k has high posterior support. It is for this reason that not too vague priors were chosen for β_k $k = 1, \ldots, p$.

Nevertheless, an alternative formulation that attempts to circumvent the problem sampling β_k from a too vague prior when $\gamma_k = 0$ is available due to Dellaportas *et al* (1997) and is described in the next subsection.

2.3.2 Gibbs Variable Selection

This method is a variation of the method of Kuo and Mallick and is aimed at circumventing the problem of sampling β_k from a too vague prior. The linear predictor is specified exactly as done in the method of Kuo and Mallick (4) but this time when $\gamma_k = 0$, β_k is sampled from the so-called pseudo-prior which is a prior distribution that has no impact on the posterior distribution. The pseudo-prior serves the purpose of increasing the efficiency of the sampler by allowing the sampler to propose good values for β_k conditional on the value of γ_k (0 or 1) (Carlin and Chib, 1995; O'Hara and Sillanpaa, 2009). Specification of the pseudo-prior is facilitated by the so called spike and slab priors (Mitchell

and Beauchamp, 1988).

The idea behind spike and slab priors is to think of $\beta_k s$ as arising from a mixture of two normal distributions which have different variances. This has an important implication when $\gamma_k = 0$ since β_k will no longer be sampled from its prior distribution as in the method of Kuo and Mallick. Essentially, this specification of the prior translates to the following spike and slab prior for β_k:

$$\beta_k|\gamma_k \sim (1-\gamma_k)N(\mu_k, \tau_{0k}^2) + \gamma_k N(0, \tau_{1k}^2)$$

where $\tau_{0k}^2 << \tau_{1k}^2$. The implication of this specification is that β_k is dependent on γ_k a priori, i.e. $f(\beta_k, \gamma_k) = f(\beta_k|\gamma_k)f(\gamma_k)$.

For the choice of pseudo-prior parameters (μ_k and τ_{0k}^2) we were guided by Dellaportas *et al* (1997) who suggest that a possible choice is to take these from a pilot run of the full model, i.e. model with all covariates included. We also considered arbitrary choices for μ_k and τ_{0k}^2 to demonstrate that one needs to exercise caution with selection of these values. As for the slab variance we took $\tau_{1k}^2 = 100$ to reflect no prior information. As in subsection 2.3.1, the other priors are chosen to reflect little prior information as

i. $\gamma_k \sim Bern(p_k)$ with $p_k \sim U(0,1)$ for $k = 1, 2, \ldots, p$;

ii. $\beta_0 \sim dflat()$; and

iii. IG$(0.001, 0.001)$ for each of σ^2, σ_u^2 and σ_v^2.

Note that in this model the Gibbs sampler full conditionals for β_k are obtained as:

$$f(\beta_k|\boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{\beta}_{(k)}) \propto \begin{cases} f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})N(0, \tau_{1k}^2) & \text{if } \gamma_k = 1, \\ N(\mu_k, \tau_{0k}^2) & \text{if } \gamma_k = 0. \end{cases}$$

2.3.3 Stochastic Search Variable Selection

This method differs from Kuo and Mallick and GVS in that $\boldsymbol{\beta}$ is specified to be of full dimension p under all models, so that the linear predictor is

$$\eta_i^{(j)} = \sum_{k=1}^{p} \beta_k X_{ki}^{(j)} \tag{5}$$

as in model (1). Indicator variables γ_k come in via the specification of the prior of β_k as follows:

$$\beta_k|\gamma_k \sim (1-\gamma_k)N(0,\tau_k^2) + \gamma_k N(0, c_k^2\tau_k^2) \tag{6}$$

and

$$P(\gamma_k = 1) = 1 - P(\gamma_k = 0) = p_k. \tag{7}$$

When $\gamma_k = 0$ (i.e. covariate X_k is excluded), it is assumed that β_k has a spike prior, i.e. $\beta_k \sim N(0,\tau_k^2)$ and when $\gamma_k = 1$ (i.e. covariate X_k is included), it is assumed that β_k has a slab prior, i.e. $\beta_k \sim N(0, c_k^2\tau_k^2)$. The principle behind this formulation is to set τ_k very small (>0) so that if $\gamma_k = 0$, then β_k would be close to zero and therefore unimportant. On the other hand, c_k is set large (c_k >1 always) to have a diffuse prior and a non-zero estimate of β_k when $\gamma_k = 1$, i.e. when it is included. As before, the parameter p_k can be thought of as the prior probability that predictor variable X_k will be included in the model i.e., $\beta_k \neq 0$ (George and McCulloch, 1993). Figure 1 depicts the spike and slab priors.

George and McCulloch (1993) suggested that the choice of c_k must allow for a non-zero estimate of β_k in the model whenever $\beta_k \sim N(0, c_k^2\tau_k^2)$. On the other hand, they suggest choosing c_k large enough to allow for values of β_k that are considerably different from zero, but rather not too large to support unrealistic values of β_k. They note that the two normal distributions in Figure 1 intersect at the points $\pm\varepsilon_k = \tau_k\delta_k$ where

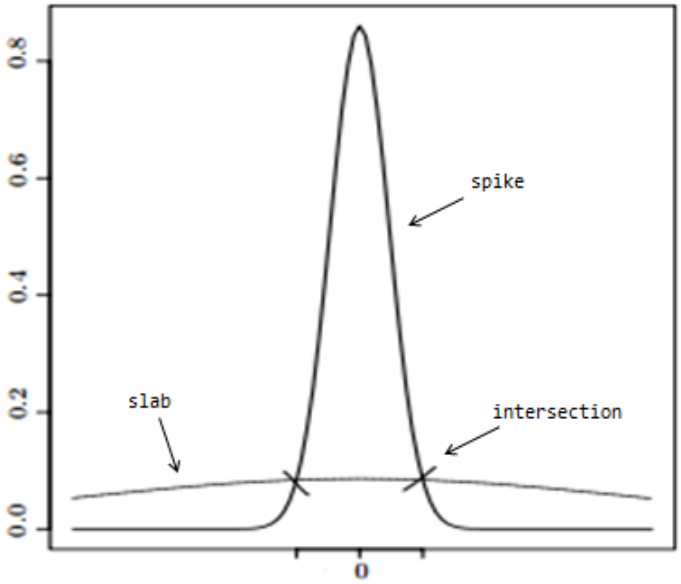

Figure 1: Spike and slab priors of George and McCulloch (1993)

$\delta_k = \sqrt{2(\log c_k)c_k^2/(c_k^2 - 1)}$. The point ε_k can be taken as a margin of 'statistical significance' such that regression coefficients falling into the interval $[-\varepsilon_k,\varepsilon_k]$ can be interpreted to be zero. Therefore, the tuning parameters can be chosen in such a way that the point of intersection reflects one's perception of statistical significance. Alternatively, τ_k can be assumed unknown and therefore estimated from the data, with this assumption only the value of c_k is fixed (Meuwissen and Goddard, 2004).

It is well documented in the literature that the choice of the tuning parameters heavily influences posterior model probabilities (and of course, posterior inclusion probabilites of covariates), see among others George and McCulloch (1997); Dellaportas *et al* (2000) and O'Hara and Sillanpaa, (2009). In view of this, several choices of the tuning parameters were considered as a sensitivity analysis.

Again, as in subsection 2.3.1 the other priors are chosen to reflect little prior information as:

i. $\gamma_k \sim Bern(p_k)$ with $p_k \sim Beta(1,1)$ for $k = 1, 2, \ldots, p$;

ii. $\beta_0 \sim dflat()$; and

iii. IG$(0.001, 0.001)$ for each of σ_u^2 and σ_v^2.

and the Gibbs sampler full conditionals for β_k are obtained as:

$$f(\beta_k|\boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{\beta}_{(k)}) \propto f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) f(\beta_k|\gamma_k)$$

and thus values of the prior when $\gamma_k = 0$ influence the posterior distribution.

2.4 Bayesian Variable Selection with Spatially Varying Coefficients Model

Model (1) assumes that regression coefficients are fixed over the entire region. While this is indeed a reasonable assumption it can be argued that the regression coefficients are not fixed but rather vary from area to area. Gelfand *et al* (2003) introduced a model which assumes spatially varying regression coefficients. Extending their model to our setting, model (1) becomes:

$$\begin{aligned} Y_i^{(j)} &= \beta_0^{(j)} + \beta_1^{(j)} X_{1i}^{(j)} + \cdots + \beta_p^{(j)} X_{pi}^{(j)} + \varepsilon_i^{(j)} \\ &= \beta_0^{(j)} + \sum_{k=1}^{p} \beta_k^{(j)} X_{ki}^{(j)} + \varepsilon_i^{(j)} \end{aligned} \tag{8}$$

where $\beta_0^{(j)}, \beta_1^{(j)}, \ldots, \beta_p^{(j)}$ are area specific regression coefficients and $\varepsilon_i^{(j)} \sim N(0, \sigma^2)$ is the error term for subject i in area j. At this stage our interest is to combine model (8) with BVS techniques described in subsection 2.3. By incorporating latent covariate inclusion indicators at area level the resulting model allows identification of important predictor variables at area level - this way each area model includes different covariates.

This type of modelling has few references in the literature and here we are guided by the work of Lum (2012). In their paper, they employ spike and slab priors for $\beta_k^{(j)}$ similar to

those used by George and McCulloch (1993) in their SSVS approach (subsection 2.3.3). They assume that,

$$\beta_k^{(j)} \sim \gamma_k^{(j)} N(\mu_k, \tau_k^2) + (1 - \gamma_k^{(j)})\delta_0(\beta_k^{(j)}) \tag{9}$$

where $\delta_0(\beta_k^{(j)})$ is the Dirac delta function evaluated at $\beta_k^{(j)}$. The area specific covariate inclusion indicators are modelled using a probit model with a spatial dependence structure (via a CAR prior) so as to achieve spatial information sharing in variable selection as follows,

$$\text{probit}(\gamma_k^{(j)}) = \eta_{0k} + \eta_k^{(j)} \tag{10}$$

where η_{0k} is a constant term and $\eta_k^{(j)} \sim$ CAR prior. This prior induces spatially smoothed inclusion probabilities such that if a covariate is included among neighbours of area j then this covariate has high chance of being included for area j. To complete their model they place a N(.,.) prior on the μ_ks and an IG$(\varepsilon, \varepsilon)$ on the τ_k^2s.

In this project we took a more or less similar approach to that of Lum (2012). For regression coefficients we placed common priors on the $\beta_k^{(j)}s$. As for the entry parameters $(\gamma_k^{(j)})$ we took three approaches.

As a starting point, we assumed that the entry parameters are inexchangeable such that in estimating them, no information is shared among the areas. In other words, we assumed that the importance of a given predictor variable in a given area has no similarity with its importance in any other area i.e.,

$$\gamma_k^{(j)} \sim Bern(p_k^{(j)}) \tag{11}$$

where $p_k^{(j)} \sim Beta(1, 1)$. This assumption may however be too strong to hold in a spatial setting. In the context of our data, it helps to note that cultural behaviors (e.g. dietary habits) as well as other aspects such as iodine content in the soil do not change in accor-

dance with administrative boundaries, and therefore, it may as well be more sensible to allow for information sharing in the entry parameters.

In the second approach the idea was to share information among neighbours using an approach similar to Lum (2012) as follows,

$$\text{logit}(\gamma_k^{(j)}) = \eta_{0k} + \eta_k^{(j)} \tag{12}$$

where η_{0k} is a constant term and $\eta_k^{(j)} \sim$ CAR.

In the third approach the idea was to share information across the entire region by modelling the entry parameters in terms of a normal prior as follows,

$$\text{logit}(\gamma_k^{(j)}) = \eta_{0k} + \eta_k^{(j)} \tag{13}$$

where η_{0k} is a constant term and $\eta_k^{(j)} \sim \text{N}(.,.)$.

In terms of variable selection, we employed the method of Kuo and Mallick and SSVS. We did not consider GVS in this setting as it requires specification of a lot of parameters for the pseudo-priors - in this case as many pseudo-priors as the number of areas would have to be specified.

2.4.1 Kuo and Mallick

Employing the method of Kuo and Mallick, the linear predictor of model (8) becomes $\eta_i^{(j)} = \sum_{k=1}^{p} \gamma_k^{(j)} \beta_k^{(j)} X_{ki}^{(j)}$. Prior distributions are specified to reflect no prior information as follows:

i. $\beta_k^{(j)} \sim N(\mu_k, \tau_k^2)$ for $k = 0, 1, \ldots, p$ with $\mu_k \sim N(0, 10)$ and $\tau_k^2 \sim \text{IG}(0.001, 0.001)$;

ii. $\text{IG}(0.001, 0.001)$ for σ^2; and

iii. In Model A: $\gamma_k^{(j)} \sim Bern(p_k^{(j)})$ with $p_k^{(j)} \sim Beta(1, 1)$. In Model B: $\text{logit}(\gamma_k^{(j)}) =$

$\eta_{0k}+\eta_k^{(j)}$ where $\eta_k^{(j)} \sim$ CAR. In Model C: $\text{logit}(\gamma_k^{(j)}) = \eta_{0k}+\eta_k^{(j)}$ where $\eta_k^{(j)} \sim \text{N}(0, 10)$ for $k = 1, 2, \ldots, p$.

2.4.2 Stochastic Search Variable Selection

As for the SSVS method, the linear predictor of model (8) is retained as $\eta_i^{(j)} = \sum_{k=1}^{p} \beta_k^{(j)} X_{ki}^{(j)}$. Prior distributions are specified to reflect no prior information as follows:

i. $\beta_k^{(j)} \sim \gamma_k^{(j)} N(\mu_k, \tau_k^2)+(1-\gamma_k^{(j)})N(0, c_k^2\tau_k^2)$ with $\mu_k \sim N(0, 10)$ and $\tau_k^2 \sim \text{IG}(0.001, 0.001)$;

ii. $\text{IG}(0.001, 0.001)$ for σ^2; and

iii. In Model A: $\gamma_k^{(j)} \sim Bern(p_k^{(j)})$ with $p_k^{(j)} \sim Beta(1, 1)$. In Model B: $\text{logit}(\gamma_k^{(j)}) = \eta_{0k}+\eta_k^{(j)}$ where $\eta_k^{(j)} \sim$ CAR. In Model C: $\text{logit}(\gamma_k^{(j)}) = \eta_{0k}+\eta_k^{(j)}$ where $\eta_k^{(j)} \sim \text{N}(0, 10)$ for $k = 1, 2, \ldots, p$.

Note that in (i.) we can no longer fix c and τ_k as was done in subsection 2.3.3, this is because the spike and the slab are now centered at different points, i.e. at zero and μ_k respectively. We therefore followed the random effects variant of SSVS due to Meuwissen and Goddard (2004) where τ_k is taken as a parameter to be estimated and c can be fixed by the user.

In terms of determining which covariate is included or excluded for any given area, we employed the MP model approach of Barbieri and Berger (2004).

3 Results

In this section study results are presented. First, we present some exploratory analysis that was conducted to get a feel of the data. This is followed by results of variable selection in the case of a spatially fixed coefficients model and lastly variable selection in the case of a spatially varying coefficients model.

3.1 Exploratory Data Analysis

Three exploratory analyses are presented here. We give an overview in terms of sample sizes in the municipalities, the spread of the response variable and the correlation among the predictor variables.

Of the 315 municipalities of Galicia, 289 were used for analysis after removing those which had no data as well as observations with missing predictor variables and/or TSH values. Figure 2 provides a summary of the number of observations in the 289 municipalities in the year of 2009. From this map we learn that there is a significant number of municipalities with few data, 97 have between 1 and 5 ([1,5)) and 50 have between 5 and 10 ([5,10)). It is useful to keep this in mind when it comes to modelling, particularly for the spatially varying coefficients model, since some areas do not have sufficient data to tell whether or not a predictor variable is important.

Figure 3 shows a histogram of the response variable (TSH) as well as that of a transformed version of it ($\log$ TSH). It can be seen that the response variable is highly skewed, hence the need to consider a transformation. In model (1) we assume a normal distribution for the response and it therefore makes sense to consider a fairly symmetrical version of TSH.

In regression analysis problems that arise due to collinearity among predictor variables are well documented (see for example Kutner *et al*, 2005). We therefore checked for the presence of multicollinearity by calculating pairwise correlations amongst the four

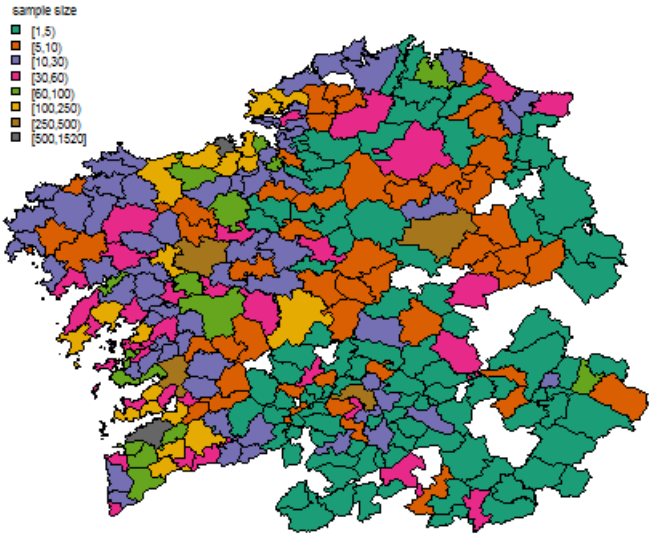

Figure 2: Frequency of observations

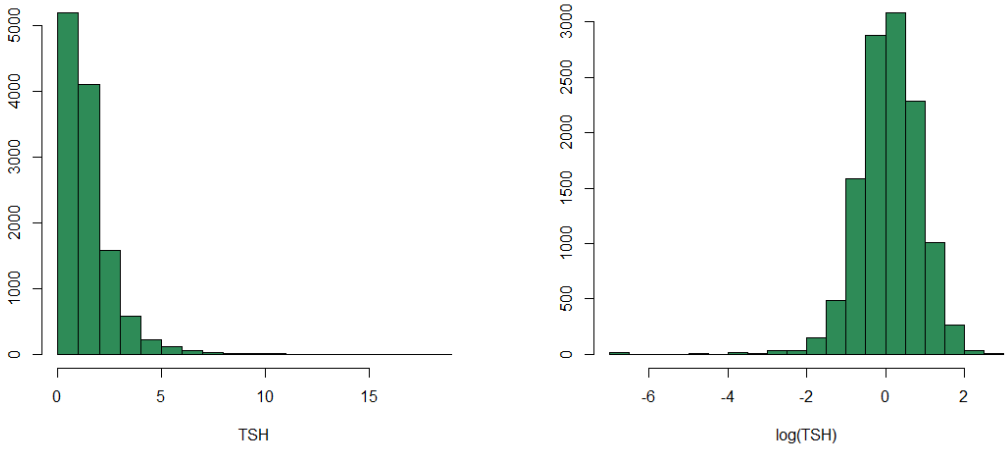

Figure 3: Histograms: TSH and log(TSH)

predictor variables. We note that the correlations are not too high to require remedial action (Table 1). Predictor variables were however centered for computational reasons.

Table 1: Correlation matrix of predictor variables

Variable	gestage	bthwgt	timetofeed
gestage	1		
bthwgt	0.324	1	
timetofeed	0.006	0.004	1

3.2 Variable Selection: spatially fixed coefficients setting

In this section we present variable selection results based on model (1) - here, the number of predictor variables is taken to be four, i.e. $p = 4$. We begin by presenting results of classical variable selection using DIC followed by those of BVS.

3.2.1 Deviance Information Criterion

Since the number of predictor variables is four, the total number of possible subsets of predictors is $2^4 = 16$. Table 2 is a listing of these subsets together with their DIC values. The 'best' subset is the one that yields the least DIC, which in this case is model (2). However, it can be noted that there are other subsets which are almost as good, i.e. those that differ from the 'best' subset by a difference less than 5 (models 1, 5 and 9). A sensible choice is then to select the smallest subset among these, i.e. model including birth weight and gender.

Table 2: Possible models and their DIC values

subset	gestage	bthwgt	gender	timetofeed	DIC
1	⊗	⊗	⊗	⊗	**27632.0**
2	⊗	⊗	⊗	-	**27631.3**
3	⊗	⊗	-	⊗	27641.9
4	⊗	-	⊗	⊗	27651.5
5	-	⊗	⊗	⊗	**27635.0**
6	⊗	⊗	-	-	27640.7
7	⊗	-	⊗	-	27650.8
8	⊗	-	-	⊗	27657.5
9	-	⊗	⊗	-	**27634.3**
10	-	⊗	-	⊗	27644.3
11	-	-	⊗	⊗	27650.5
12	⊗	-	-	-	27656.6
13	-	⊗	-	-	27643.4
14	-	-	⊗	-	27649.7
15	-	-	-	⊗	27656.3
16	-	-	-	-	27655.2

⊗ means present, - means absent

3.2.2 Bayesian Variable Selection

Implementation of the method of Kuo and Mallick is straight forward in that it does not involve choosing tuning parameters, however the GVS and SSVS techniques do (subsection 2.3). For the GVS technique, values for the pseudo-prior parameters were obtained from a pilot run of the full model as suggested by Dellaportas *et al* (1997). They suggest

use of full model parameter estimates and standard errors as choices for μ_k and τ_{0k}, we therefore took (0.0650,-0.0800,0.0500,-0.0150) and (0.0280,0.0170,0.0142,0.0130) respectively. To demonstrate that one needs to exercise caution with selection of these values we took $\mu_k = (0, 0, 0, 0)$ and $\tau_{0k} = (0.0140, 0.0085, 0.0071, 0.0065)$.

As for the SSVS technique, different settings were specified for sensitivity analysis purposes. We observed the magnitude of the regression coefficients and thereof assumed three margins of 'statistical significance' namely, [-0.01,0.01], [-0.0175,0.0175] and [-0.025,0.025]. In terms of the choice of c, i.e. the factor by which the spike standard deviation is multiplied to get a larger standard deviation for the slab, the values were taken as $c = 100$, $c = 125$ and $c = 100$ respectively.

For all methods we ran 10,000 overrelaxed iterations with thinning equal to 3, furthermore, a burn-in of 1,000 was removed. In terms of speed, we noted that SSVS was the fastest, followed by method of Kuo and Mallick and lastly GVS.

Table 3 shows the posterior inclusion probabilities of each predictor variable based on the method of Kuo and Mallick, GVS and SSVS variable selection techniques. By employing the MP model principle (Barbieri and Berger, 2004), we are led to select the predictors birth weight and gender as the important ones to include in the regression model.

Table 3: Posterior inclusion probabilities for predictor variables

method	tuning parameters	inclusion probability				speed (in sec.)
		gestage	bthwgt	gender	timetofeed	
K & M	-	0.10	**0.92**	**0.50**	0.01	2496
GVS	$\mu_k = (0.0650, -0.0800, 0.0500, -0.0150)$ $\tau_{0k} = (0.0280, 0.0170, 0.0142, 0.0130)$	0.09	**0.89**	**0.50**	0.01	2685
	$\mu_k = (0.0000, 0.0000, 0.0000, 0.000)$ $\tau_{0k} = (0.0140, 0.0085, 0.0071, 0.0065)$	0.07	**0.85**	0.1	0.01	2691
SSVS	$c = 100$, $\varepsilon = 0.0100$, $\tau_k = 0.0033$	0.44	**0.98**	**0.50**	0.10	581
	$c = 125$, $\varepsilon = 0.0175$, $\tau_k = 0.0056$	0.30	**0.95**	0.45	0.038	493
	$c = 100$, $\varepsilon = 0.0250$, $\tau_k = 0.0082$	0.23	**0.92**	**0.50**	0.03	504

Table 4 shows posterior model probabilities also based on the method of Kuo and Mallick, GVS and SSVS variable selection techniques. The acronyms GVS_a, GVS_b and SSVS_a,

$SSVS_b$, $SSVS_c$ denote the different tuning parameter settings for GVS and SSVS respectively. By employing the MAP principle, we are led to select the predictor variable birth weight as the important one to include in the regression model. Though all the three approaches favour the model with birth weight only, we observe that the model with birth weight and gender also appears frequently in the MCMC sample - it appears more in the method of Kuo and Mallick followed by GVS and lastly SSVS.

Table 4: Posterior model probabilities

subset	gestage	bthwgt	gender	timetofeed	K & M	GVS_a	GVS_b	$SSVS_a$	$SSVS_b$	$SSVS_c$
1	⊗	⊗	⊗	⊗	0.0004	0.0008	0.0000	0.0147	0.0059	0.0026
2	⊗	⊗	⊗	-	0.0630	0.0497	0.0083	0.2054	0.1412	0.1091
3	⊗	⊗	-	⊗	0.0003	0.0004	0.0000	0.0197	0.0053	0.0033
4	⊗	-	⊗	⊗	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	-	⊗	⊗	⊗	0.0029	0.0040	0.0007	0.0208	0.0102	0.0118
6	⊗	⊗	-	-	0.0412	0.0344	0.0566	0.1982	0.1466	0.1084
7	⊗	-	⊗	-	0.0000	0.0004	0.0000	0.0001	0.0007	0.0020
8	⊗	-	-	⊗	0.0000	0.0000	0.0000	0.0000	0.0002	0.0010
9	-	⊗	⊗	-	**0.4114**	**0.3697**	0.0788	**0.2274**	**0.2757**	**0.3080**
10	-	⊗	-	⊗	0.0019	0.0031	0.0067	0.0206	0.0101	0.0109
11	-	-	⊗	⊗	0.0000	0.0040	0.0000	0.0008	0.0003	0.0004
12	⊗	-	-	-	0.0004	0.0008	0.0016	0.0010	0.0029	0.0034
13	-	⊗	-	-	**0.4152**	**0.4247**	**0.6954**	**0.2758**	**0.3482**	**0.3538**
14	-	-	⊗	-	0.0078	0.0249	0.0072	0.0053	0.0162	0.0263
15	-	-	-	⊗	0.0001	0.0001	0.0009	0.0007	0.0012	0.0014
16	-	-	-	-	0.0552	0.0870	0.1439	0.0096	0.0352	0.0583

⊗ means present, - means absent

Thus, we note that in general, the MP and MAP principles do not lead to the selection of the same model. To choose which principle to adopt, it helps to note that in this case the MP model coincides with a good predictive model (as identified by the DIC). These findings are in line with Barbieri and Berger (2004) who showed that oftentimes the high probability model (i.e. MAP model) is not optimally predictive but rather, the median model (i.e. MP model) is.

It can also be noted that the choice of tuning parameters in the SSVS approach affects the posterior probabilities - the predictor gender would appear unimportant in setting $SSVS_b$. This is not surprising as it is well documented in the literature that the choice of tuning parameters can drive the results (see George and McCulloch 1993; George and McCulloch 1997; Dellaportas *et al*, 1997; Kuo and Mallick, 1998; O'Hara and Sillanpaa,

2009; Lesaffre and Lawson, 2012). Moreover, we note that an arbitrary choice of pseudo-prior parameters for the GVS method will not give good results (GVS_b), other possible choices are given in Ntzoufras (1999).

Summarizing, we choose birth weight and gender as the important predictor variables to include in the regression model. The choice is in accordance with results from classical variable selection (based on DIC) and BVS (used with MP principle). Though these two techniques agree to a large extent, we note that BVS give the benefit of avoiding fitting of all 16 possible models. Parameter estimates of the final model are shown in Table 5.

Table 5: Parameter estimates of final model

Parameter	Estimate (std. error)
intercept	0.142 (0.016)
bthwgt	-0.067 (0.014)
gender	0.047 (0.022)
σ^2	0.600 (0.008)
σ_u^2	0.006 (0.003)
σ_v^2	0.035 (0.01)

3.3 Variable Selection: spatially varying coefficients setting

In this subsection we present results of BVS applied to the spatially varying coefficients model (8). Before presenting the main results we give an insightfull preliminary analysis.

3.3.1 Preliminary results

Since the aim is to extend variable selection to a spatially varying coefficients model, we began by fitting this model with no variable selection - this was done to get a rough idea about the importance of each of the four predictor variables in the different areas. Two choices of priors were considered for the regression coefficients namely $\beta_k^{(j)} \sim \text{N}(\mu_k, \sigma_k^2)$ with $\theta_k \sim \text{N}(0, 100)$ and $\sigma_k^2 \sim \text{IG}(0.001, 0.001)$ as well as $\beta_k^{(j)} \sim$ CAR prior.

Figure 4 depicts a summary of parameter estimates of $\beta_k^{(j)}s$ based on the normal prior (results obtained from assuming a CAR prior are shown in the Appendix). The standard

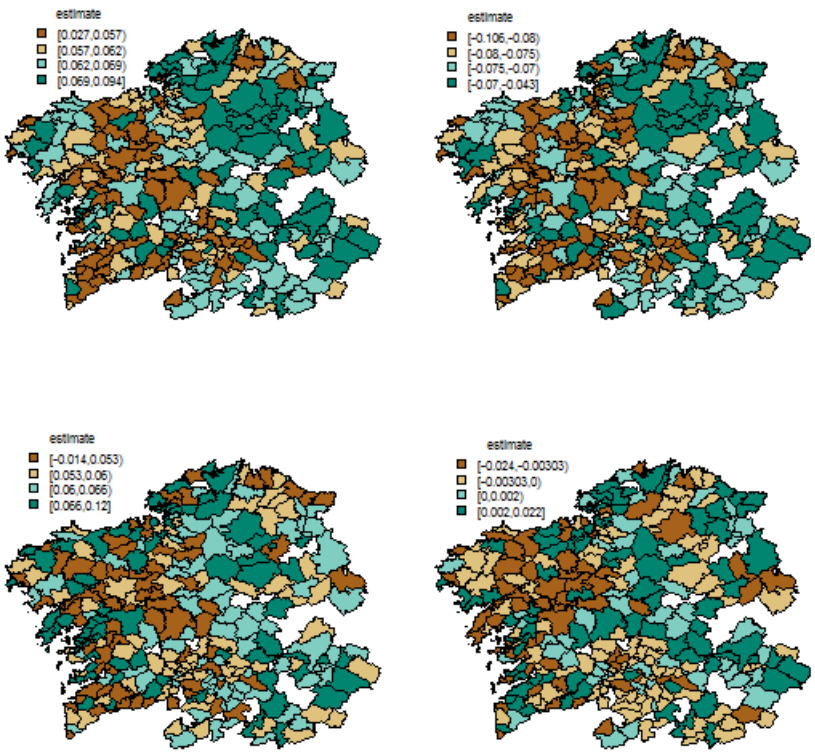

Figure 4: Summary: parameter estimates of the four predictor variables

errors range between 0.028-0.034, 0.030-0.042, 0.040-0.080 and 0.003-0.023 for gestational age, birth weight, gender and time to feed respectively. We observe that the $\beta_k^{(j)}$ estimates are shrunk towards their respective means (Table 6), especially for municipalities with few data, moreover, we also observe that on average standard errors are large relative to the parameter estimates. Based on this preliminary analysis, it would appear that at the area level few to no predictor variables are important in explaining the response variable.

Table 6: Estimates of regression coefficient prior means

Predictor	Estimate (std. error)
gestage	0.063 (0.026)
bthwgt	-0.075 (0.018)
gender	0.060 (0.018)
timetofeed	-0.0006 (0.003)

3.3.2 Main results

We now present results of BVS. The method of Kuo and Mallick and two settings of SSVS were considered (taking $c^2 = 10, c^2 = 100$ denoted SSVS_a, SSVS_b respectively). However, here we only show results of the method of Kuo and Mallick since all results lead to more or less the same conclusions, SSVS results are shown in the Appendix. Note that we only have results of Model A (where it is assumed that entry parameters are inexchangeable) since Models B and C did not reach convergence.

Figure 5 depicts Model A posterior inclusion probabilities (per predictor variable). An inclusion probability greater or equal to 0.5 is taken to mean that a given predictor variable is important in a given area. We observe that predictor variable inclusion appears to occur mostly at locations with few data. The explanation to this counter-intuitive finding is that in such areas there is little information to update the prior distribution and as a result the average inclusion probability of prior distribution (which is equal to 0.5) dominates the likelihood and hence the inclusion. This problem can however be solved if one adopts Models B and C - by assuming that the entry parameters are exchangeable the priors of these models induce information sharing such that for areas with few data entry parameters are made to appear like others on average.

Table 7 summarizes the selected predictor variables in 142 municipalities. A total of 147 municipalities with fewer than 10 observations were excluded because for such areas with few data the posterior inclusion probabilities are largely due to the prior mean. Across all three scenarios considered i.e. Kuo and Mallick, SSVS_a and SSVS_b, it is observed that in general one of the following occurs: no covariate is included, one of the four is included, birth weight and gender are included or gestational age and gender are included. Comparing with results from subsection 3.2.2, we note that the predictor variables selected in the spatially fixed coefficients model (birth weight and gender) are generally not the same as those selected at the area level.

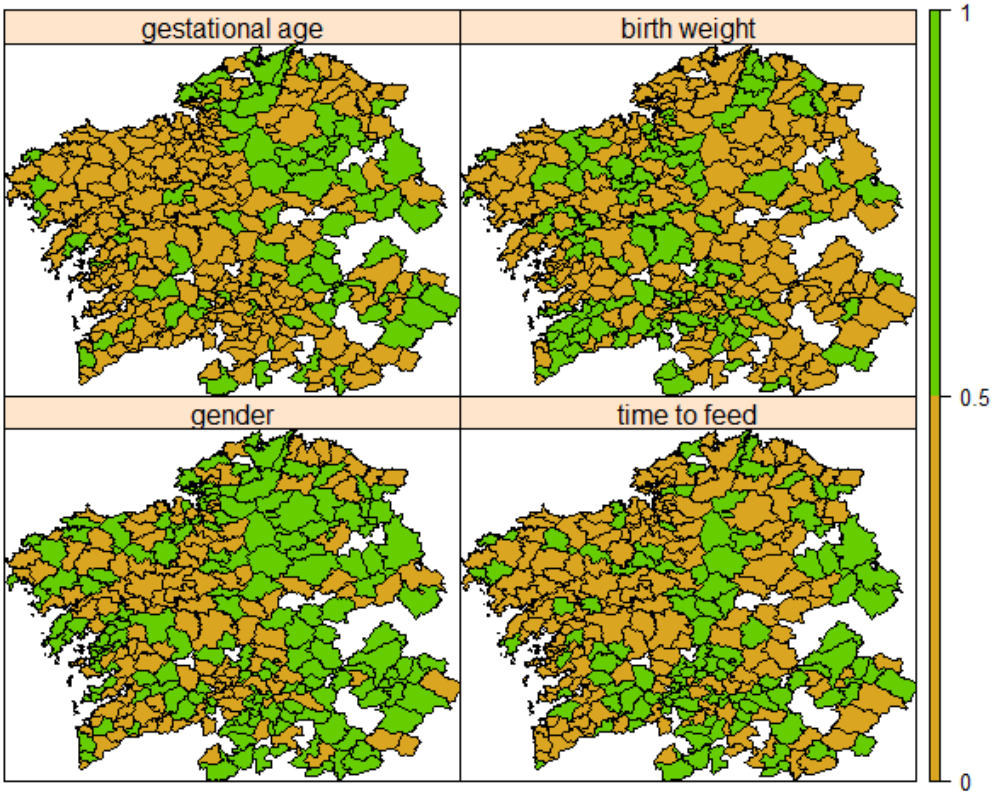

Figure 5: Variable selection at area level: Model A, Kuo and Mallick

Table 7: Frequency distribution of selected predictor variables - count (%)

subset	gestage	bthwgt	gender	timetofeed	K & M	$SSVS_a$	$SSVS_b$
1	⊗	⊗	⊗	⊗	1 (0.007)	1 (0.007)	1 (0.007)
2	⊗	⊗	⊗	-	4 (0.028)	3 (0.021)	4 (0.028)
3	⊗	⊗	-	⊗	-	-	-
4	⊗	-	⊗	⊗	4 (0.028)	6 (0.042)	3 (0.021)
5	-	⊗	⊗	⊗	3 (0.021)	3 (0.021)	2 (0.014)
6	⊗	⊗	-	-	3 (0.021)	1 (0.007)	3 (0.021)
7	⊗	-	⊗	-	15 (0.106)	13 (0.092)	16 (0.113)
8	⊗	-	-	⊗	-	-	1 (0.007)
9	-	⊗	⊗	-	10 (0.070)	12 (0.085)	9 (0.063)
10	-	⊗	-	⊗	5 (0.035)	3 (0.021)	5 (0.035)
11	-	-	⊗	⊗	4 (0.028)	7 (0.049)	5 (0.035)
12	⊗	-	-	-	4 (0.028)	3 (0.021)	3 (0.021)
13	-	⊗	-	-	21 (0.148)	27 (0.190)	23 (0.162)
14	-	-	⊗	-	28 (0.197)	24 (0.169)	26 (0.183)
15	-	-	-	⊗	12 (0.085)	12 (0.085)	9 (0.063)
16	-	-	-	-	28 (0.197)	27 (0.190)	32 (0.225)
⊗ means present, - means absent					**142 (100)**	**142 (100)**	**142 (100)**

4 Discussion

In this section we discuss the application of BVS in spatial regression models. We begin with the spatially fixed coefficients setting and thereafter, the spatially varying coefficients setting.

Spatially fixed coefficients setting

In subsection 3.2.2 we applied BVS to identify important predictor variables to be included in the regression model (model 1). It was found that birth weight and gender are important predictors of TSH levels in newborns. The results show that keeping gender constant, the higher the birth weight the lower the TSH level and keeping birth weight constant, male infants tend to have higher TSH levels compared to female infants (Table 5). These results are consistent with findings from other studies, e.g., Frank *et al* (1996) and Herbstman *et al* (2008) reported that birthweight is predictive of TSH level and Sullivan *et al* (1997), Chan *et al* (2001) and again Herbstman *et al* (2008) reported that gender is predictive of TSH level. Note that other studies have also found gestational age as predictive of TSH level (see Herbstman *et al* (2008)).

Turning to the variable selection exercise, we saw that BVS techniques require less computational effort as compared to classical approaches - with BVS one does not need to fit all the possible models in order to identify important predictor variables. However, when using BVS one needs to pay attention to the decision rule they adopt in order to select predictor variables that will yield an optimally predictive model. It was noted that the best subset of predictor variables as identified by the MAP model principle was different from that identified by the MP model principle. We went in favour of the latter as it coincided with an optimally predictive model identified using the DIC. On this note, we can mention that our findings are in agreement with Barbieri and Berger (2004) who noted that often times the MP model is optimally predictive and the MAP model is not necessarily so.

Adopting the MP model principle as the decision rule, we were led to select the same predictor variables across all three approaches, i.e. method of Kuo and Mallick, GVS and SSVS. Though there was agreement among all the approaches, we noted some advantages and disadvantages for each. We now reflect on these in turn.

An advantage of the method of Kuo and Mallick is that it has a straight forward implementation. Unlike the GVS and the SSVS, it does not require sophisticated choice of tuning parameters, one only has to specify priors for regression coefficients and entry parameters independently. When specifying priors for regression coefficients, one has to set the prior variance considerably low in order to have good mixing (see O'Hara and Sillanpaa, 2009). In terms of disadvantages we observed that this method is slow (Table 3). Other authors have also noted that this method does not work well in the presence of multicollinearity (see Ntzoufras, 1999), in our case this was not experienced since our predictors were not highly correlated.

As in the method of Kuo and Mallick, the GVS approach also has a relatively simple implementation. Moreover, if a good pseudo-prior is specified the GVS leads to better mixing if compared with the former. However, a drawback of this method is specification of pseudo-priors - sophisticated choices are required. In our analysis (Table 3) we considered a choice suggested in Dellaportas *et al* (1997) of taking values obtained from a pilot-run of the full model. This choice worked reasonably well - results obtained were in agreement with those of other approaches. In order to demonstrate that pseudo-prior parameters cannot be chosen arbitrarily we took values not far off from those obtained from the pilot run, the resulting MP model was different from that identified by the other approaches (see Table 3). As in the method of Kuo and Mallick, we observed that this method is slow and other authors have also noted that this method does not work well in the presence of multicollinearity (see Ntzoufras, 1999).

Turning to the SSVS approach, the major advantage that we noted was simulation speed, it was faster compared to the other approaches (see Table 3). A well known disadvantage (that we also noted) is the method's reliance on user defined and data dependent tuning parameters. The tuning parameters heavily influence posterior model and predictor inclusion probabilities and hence a sensitivity analysis may be ideal. From the three scenarios that we prespecified, we noted that if the constant c is kept fixed, a smaller value of ε allows predictor variables to enter the model more frequently. Also, increasing both c and ε disallows predictor variables from entering the model more frequently. It should however be noted that in some instances the reliance of SSVS on tuning parameters makes the approach flexible as it can allow the user to define their own margin of practical significance (George and McCulloch, 1993).

Spatially varying coefficients setting

In subsection 3.3 we extended the method of Kuo and Mallick and SSVS to the spatially varying coefficients model. Though we could not get all results we hoped for, the few that were obtained seemingly indicated that generally, one of the following occurs: no covariate is included, one of the four is included, birth weight and gender are included or gestational age and gender are included. This would appear to contradict results from the spatially fixed coefficients setting however, it helps to note that at the municipality level there are generally few data hence, it may well be that the data are not sufficient to show importance of the predictor variables.

Turning to the variable selection exercise, basically the same advantages and disadvantages that were observed in the spatially fixed coeffients model also apply in this model. As one would also expect, the variable selection exercise became more computationally involved and hence more time consuming - as before the SSVS settings considered were faster than the method of Kuo and Mallick (1147, 1111 and 3507 seconds respectively).

Though extending BVS to this model was possible, we were only able to achieve conver-

gence in the simplest case where we assumed that the entry parameters $\gamma_k^{(j)}$ have nothing in common, here they were not modelled but only assigned priors. Note that with a binary response, even this simplest case could not reach convergence.

We experienced great difficulties when we sought to model these entry parameters to enable information sharing among neighbours as well as across regions (Models B and C, subsections 2.4.1 and 2.4.2). Several alterations were made on the programs to try to improve convergence. Among the first alterations was standardizing the predictor variables; varying the link function on the entry parameters model (i.e. logit, probit, c log log); varying parameters on priors (e.g. on the IG(.,.) prior); changing variance priors from IG(.,.) to U(.,.) on prior standard deviations as well as adjusting tuning parameters in the SSVS.

Additional alterations involved imposing some assumptions on the model so as to reduce its complexity. In place of modelling the entry parameters we assumed that they had a common success probability, i.e. $\gamma_k^{(j)} \sim \text{Bern}(p_k)$ with $p_k \sim \text{U}(0,1)$ and also $\gamma_k^{(j)} \sim \text{Bern}(0.5)$. We also tried fixing variances on regression coefficient priors, a range of values were considered. Finally we also considered implementing the models in rjags but still no convergence was reached.

On this note we put forward that though combining spatially varying coefficient models with BVS is possible and an attractive idea, allowing for information sharing among the entry parameters increases computational intensity and hence makes the model susceptible to convergence problems. Whilst Lum (2012) succeeded in doing so, more research can be done in the future to look into ways of tackling convergence problems in this context. Note that we also succeeded sharing information in the entry parameters using a N(.,.) prior (as in Model C) on a different data set with only one predictor variable.

5 Conclusion

In conclusion we found that bayesian variable selection methods can be applied with ease in spatially fixed coefficient models and they are more advantageous compared to classical methods as they reduce computational burden of fitting all possible models. However, in spatially varying coefficient models convergence problems may occur, especially when one attempts to incorporate information sharing in entry parameters. Based on the spatially fixed coefficient model, we found that the predictor variables birth weight and gender are important in explaining TSH level in newborns. In terms of the spatially varying coefficient model and based on the results obtained, different predictor variables appear to be important at different areas.

References

- American Thyroid Association. Goitre. Available at http://www.thyroid.org/wp-content/uploads/patients/brochures/Goiter˙brochure.pdf. Extracted on 02 Sep 2014.

- Assey, V. D., Greiner, T., Mzee, R. K., Abuu, H., Mgoba, C, Kimboka, S. and Peterson, S (2006). Iodine Deficiency persists in the Zanzibar Islands of Tanzania. *Food Nutr Bull*,27, 292-299.

- Assuncao, R. M. (2003). Space-varying Coefficient Models for Small area Data. *Environmetrics*, 14, 453-473.

- Barbieri, M. M. and Berger, J. O. (2004). Optimal Predictive Model Selection. *The Annals of Statistics*, 32, 870-897.

- Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman and Hall.

- Besag, J., York, J. and Mollie , A. (1991). Bayesian Image Restoration with Two Applications in Spatial Statistics. *Annals of the Institute of Statistics and Mathematics*, 43, 159.

- Carlin, B.P. and Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society B*, 157, 473484.

- Chan, L. Y., Leung, T. N. and Lau, T. K. (2001). Influences of Perinatal Factors on Cord Blood Thyroid-Stimulating Hormone Level. *Acta Obstet Gynecol Scand*, 80, 10141018.

- Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2000). Bayesian Variable Selection using the Gibbs Sampler. In Dey, D. K., Ghosh, S. K., and Mallick, B. K. (eds.). *Generalized linear models: a Bayesian perspective*, 273-286. Marcel Dekker, Inc., New York.

- Dellaportas, P., Forster, J. J. and Ntzoufras, I. (1997). On Bayesian Model and Variable Selection Using MCMC. *Technical Report*, Department of Statistics, Athens University of Economics and Business, Athens, Greece.

- Frank, J. E., Faix, J. E., Hermos R. J., Mullaney, D. M., Rojan, D. A., Mitchell, M. L. and Klein, R. Z (1996). Thyroid Function in very Low Birth Weight Infants: Effects on Neonatal Hypothyroidism Screening. *J Pediatr*, 128, 548554.

- Gelfand, A. E., Kim, H. J., Sirmans, C. F. and Banerjee, S. (2003). Spatial Modelling with Spatially Varying Coefficient Processes. *Journal of the American Statistical Association*, 462, 387-396.

- Gelman, A. (2006). Prior Distributions for Variance Parameters in Hierarchical Models. *Bayesian Analysis*, 1, 515 - 533.

- George, E. I. and McCulloch, R. E. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, 88, 881889.

- George, E. I. and McCulloch, R. E. (1993). Approaches for Bayesian Variable Selection. *Statistica Sinica*, 7, 339-373.

- Herbstman, J., Apelberg, B. J., Witter, F. R., Panny, S. and Goldman, L. R. (2008). Maternal, Infant, and Delivery Factors Associated with Neonatal Thyroid Hormone Status. *Thyroid*, 18, 67-76.

- Ishwaran, H. and Rao, J. S. (2005). Spike and Slab Variable Selection: Frequentist and Bayesian Strategies. *The Annals of Statistics*, 33, 730-773.

- Kutner, M. H., Nachtsheim, C. J., Neter, J, and Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill.

- Lesaffre, E. and Lawson, A. B. (2012). *Bayesian Biostatistics*. John Wiley and Sons.

- Lum, K. (2012). Bayesian Variable Selection for Spatially Dependent Generalized Linear Models. Available at http://arxiv.org/pdf/1209.0661.pdf. Extracted on 27 Jul 2014.

- Meuwissen, T. H. E. and Goddard, M. E. (2004). Mapping Multiple QTL using Linkage Disequilibrium and Linkage Analysis Information and Multitrait Data." *Genet. Sel. Evol.*, 36, 261-279.

- Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83, 1023-1036.

- Ntzoufras, I. (1999). Aspects of Bayesian Model and Variable Selection Using MCMC. *Unpublished Ph.D. Thesis*, Department of Statistics, Athens University of Economics and Business, Athens, Greece.

- Ntzoufras, I. (2009). *Bayesian Modelling using WinBUGS*. John Wiley and Sons.

- Kuo, L. and Mallick, B. (1998). Variable Selection for Regression Models. *Sankhya B*, 60, 6581.

- Lope, V., Pollan, M., Perez-Gomez, B., Aragones, N., Ramis, R., Gomez-Barroso, D. and Lopez-Abente, G. (2006). Municipal mortality due to cancer in Spain. Available at http://www.biomedcentral.com/1471-2458/6/302. Extracted on 25 Aug 2014.

- O'Hara, R. B. and Sillanpaa, M. J. (2009). A Review of Bayesian Variable Selection Methods: What, How and Which. *Bayesian Analysis*, 4, 85-118.

- Reich, B. J., Fuentes, M., Herring, A. H. and Evenson, K. R. (2010). Bayesian Variable Selection for Multivariate Spatially-Varying Coefficient Regression. Bayesian variable selection for multivariate spatially varying coefficient regression. *Biometrics*, 66, 772-782.

- Sullivan, K. M., May, W., Nordenberg, D., Houston, R. and Maberly, G. F. (1997).

Use of Thyroid Stimulating Hormone Testing in Newborns to identify Iodine Deficiency. *J Nutr*, 127, 5558.

- Wakefield, J. (2006). Disease mapping and Spatial Regression with Count Data. *Biostatistics*, 8, 158-183.

- World Health Organization (2007). Assessment of Iodine Defficiency Disorders and Monitoring their Elimination: A Guide for Program Managers. *3rd ed. Technical Report*. World Health Organization (WHO) and United Nations Children's Fund (UNICEF) and International Council for the Control of Iodine Defficiency Disorders, Geneva, Switzerland (WHO/NHD/01.1).

Appendix

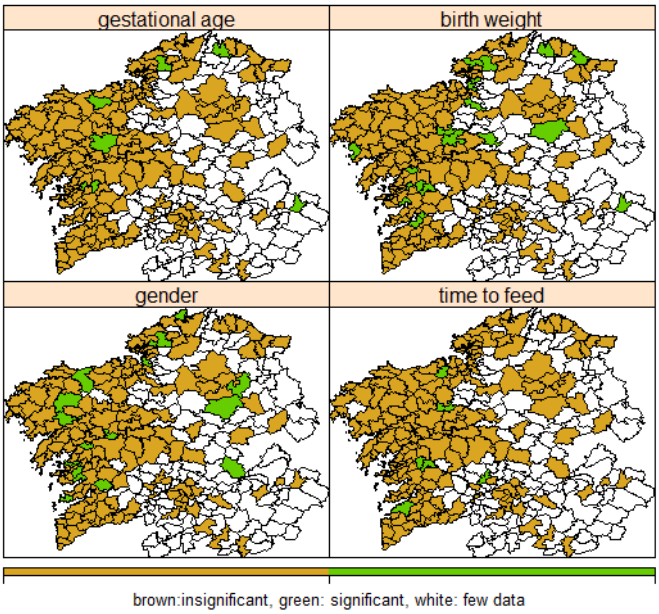

Figure 6: Summary: p-values based on OLS per municipality

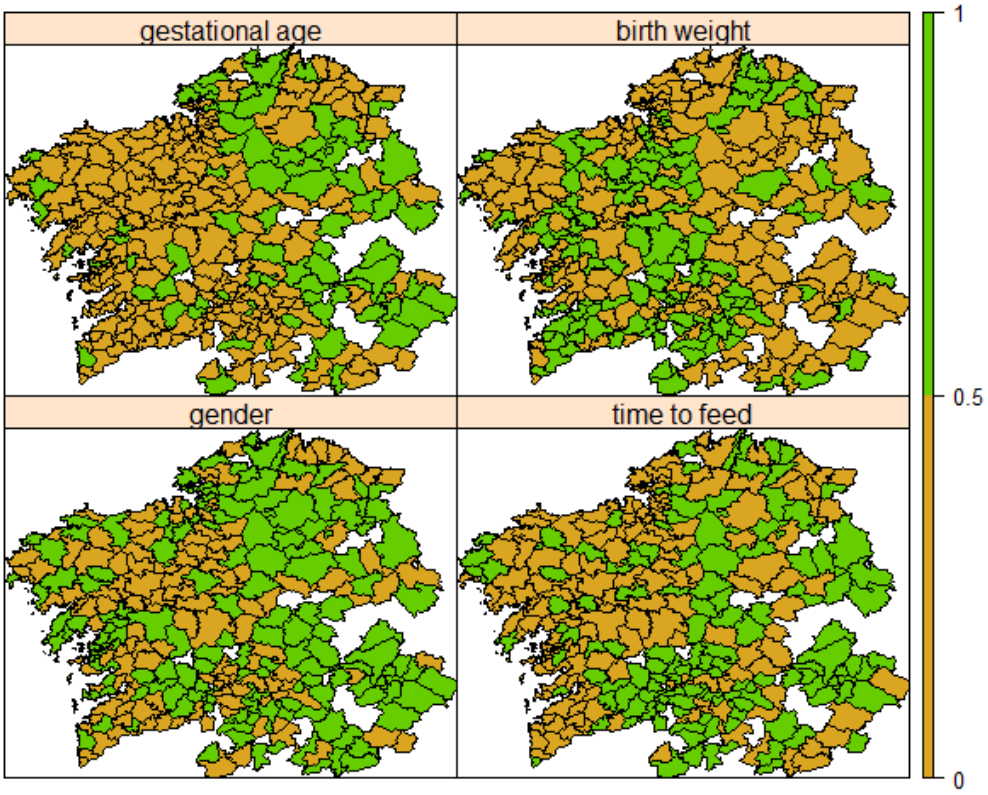

Figure 7: Variable selection at area level: Model A, SSVS with $c^2 = 10$

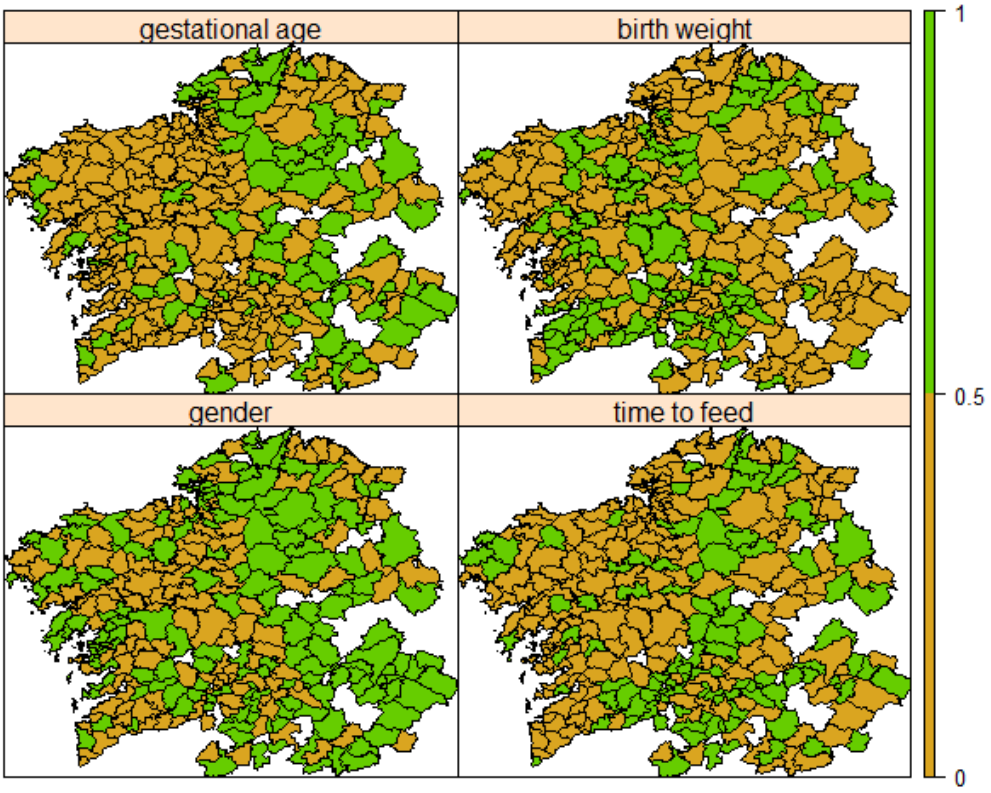

Figure 8: Variable selection at area level: Model A, SSVS with $c^2 = 100$

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Use of Bayesian variable selection methods in spatial regression models

Richting: **Master of Statistics-Biostatistics**
Jaar: **2014**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Ganyani, Tapiwa

Datum: **10/09/2014**