Journal de la Société Française de Statistique
Vol. 159 No. 2 (2018)

Bayesian selection of mixed covariates from a latent layer: application to hierarchical modeling of soil carbon dynamics

Titre: Sélection bayésienne de covariables mixtes sur la couche latente d'un modèle hiérarchique : application à la dynamique de carbone dans le sol

Rana Jreich[1,2], Christine Hatte[1], Jérôme Balesdent[3] and Éric Parent[2]

Abstract: Soil carbon is important not only to ensure food security via soil fertility, but also to potentially mitigate global warming via increasing soil carbon sequestration. There is an urgent need to understand the response of the soil carbon pool to climate change and agricultural practices. Biophysical models have been developed to study Soil Organic Matter (**SOM**) for some decades. However, there still remains considerable uncertainty about the mechanisms that affect **SOM** dynamics from the microbial level to global scales. In this paper, we propose a statistical Bayesian selection approach to study which forcing conditions influence soil carbon dynamics by looking at the depth distribution of radiocarbon content for 159 profiles under different conditions of climate (temperature, precipitation, etc.) and environment (soil type, land-use). Stochastic Search Variable Selection (**SSVS**) is here applied to latent variables in a hierarchical Bayesian model. The model describes variations of radiocarbon content as a function of depth and potential covariates such as climatic and environmental factors. **SSVS** provides a probabilistic judgment about the joint contribution of soil type, climate and land use on soil carbon dynamics. We also discuss the practical performance and limitations of **SSVS** in presence of categorical covariates and collinearity between covariates in the latent layers of the model.

Résumé : Le carbone du sol est important non seulement pour assurer la sécurité alimentaire en maintenant la fertilité des sols, mais aussi pour limiter le réchauffement climatique en augmentant la séquestration du carbone dans le sol. Il est urgent de comprendre la réaction du carbone du sol face au réchauffement climatique et au changement des pratiques agricoles. Des modèles bio-physiques ont été développés depuis quelques décennies pour étudier la matière organique du sol (**SOM**). Cependant, il existe encore une forte incertitude sur les mécanismes contrôlant la dynamique de la **SOM**, du niveau microbien aux échelles globales. Dans cet article, nous proposons une approche statistique bayésienne de sélection de variables pour mieux cerner la dynamique du carbone du sol en examinant la variation en profondeur du radiocarbone pour 159 profils sous différentes conditions de climat (température, précipitations, ...) et d'environnement (type de sol, type d'usage du sol, ...). La recherche stochastique de sélection de variables (**SSVS**) est appliquée au niveau des variables latentes d'un modèle bayésien hiérarchique. Ce modèle décrit la variation du radiocarbone en fonction de la profondeur et en tenant compte des covariables explicatives potentielles tels que les facteurs climatiques et environnementaux. Cette approche nous permet d'avoir un jugement probabiliste sur la contribution conjointe du type de sol, du climat et de l'usage du sol à la dynamique verticale du carbone dans le sol. Nous discutons également de la performance pratique et des limitations de **SSVS** en présence de covariables catégorielles et de la colinéarité entre certaines covariables quand elles interviennent au niveau d'une couche latente d'un modèle bayésien hiérarchique.

[1] Laboratoire des Sciences du climat et de l'environnement, LSCE/IPSL, UMR 8212 CEA-CNRS-UVSQ, Université Paris Saclay, F-91198 Gif-sur-Yvette, France.
E-mail: rana.jreich@lsce.ipsl.fr and E-mail: christine.hatte@lsce.ipsl.fr

[2] AgroParisTech, UMR 518 Mathématiques et Informatiques Appliquées, F-75005 Paris, France.
E-mail: eric.parent@agroparistech.fr

[3] Aix-Marseille Université, CNRS, Collège de France, IRD, INRA, CEREGE, 13545 Aix-en-Provence, France.
E-mail: jerome.balesdent@inra.fr

Keywords: Bayesian selection approach, SSVS, spike and slab prior, hierarchical Bayesian model, latent variables, organic carbon dynamics, radiocarbon
Mots-clés : méthode bayésienne de sélection de variables, recherche stochastique de sélection de covariables, modèle hiérarchique bayésien, variables latentes, dynamique du carbone organique, radiocarbone
AMS 2000 subject classifications: 35L05, 35L70

1. Introduction

A significant current issue when trying to predict our planet's future is to understand the feedback effects between climate evolution and the future soil carbon balance. Soil constitutes the largest carbon pool in interaction with atmospheric carbon, containing 2000 to 2400 Gt of organic carbon in the first meter, i.e. at least the equivalent of 250 years of current fossil carbon emissions that are estimated at 10±0.6 Gt/year (Stocker, 2014).
The stock of soil organic matter (**SOM**) has been defined as a balance between input of organic matter through vegetation and loss through microbial decomposition. A large variation in the soil organic carbon (**SOC**) stock amongst soil types and land use has been shown, ranging from 2 $\mathrm{kg/m^2}$ for arenosols to more than 10 $\mathrm{kg/m^2}$ for podzols (Batjes, 1996). Regarding land use, Martin et al. (2011) show that relationships between soil organic carbon stocks and pedo-climate depend on the type of land use and that they differ between forest and cultivated soil.

The global analyses carried out by Carvalhais et al. (2014) and He et al. (2016) point out the lack of knowledge of carbon residence time in soil and an increasing concern about the importance of climate factors in the variability of carbon storage. For instance, a temperature increase may clearly impact the activity of soil microorganisms and the subsequent organic carbon sequestration by soils. Moreover **SOM** evolution plays a key role in the $\mathbf{CO_2}$ atmospheric content since the soil is a crucial pool for $\mathbf{CO_2}$ emission or sequestration. No consensus has been reached, however, on the relative importance of the various climatic factors that affect **SOM** dynamics, such as temperature, precipitation, aridity, moisture, etc.

In fact, several questions remain unclear for soil scientists: Could soil capacity be durably increased to sequestrate more carbon by changing land use? What quantitative changes in **SOM** occur when modifying agricultural practices? Will that change the soil carbon stock/the organic matter residence time? What is the contribution of each climatic or environmental factor to soil carbon? Is the potential increment of the soil carbon stock to be considered as sustainable ? These questions highlight the importance of assessing the uncertainties as well as understanding the complex mechanisms of soil carbon dynamics. To investigate this point through data collection, in addition to soil carbon concentration, $F^{14}C$ measurements are also taken into account to describe **SOM** dynamics on the grounds that radiocarbon content can be considered as a clock that registers **SOC** residence time (Scharpenseel, 1971).

A worldwide meta-analysis of radiocarbon profiles is described in Mathieu et al. (2015). In their study, a hierarchical non linear model is designed under the frequentist paradigm with inference performed by the "Expectation-Maximization" algorithm. The radiocarbon dynamics is parameterized as a smooth function of depth with random effects taking into account potentially explanatory climatic and environmental factors. Once calibrated, the model is used for statistical prediction along various typical scenarios of (modified) forcing conditions; according to an expert interpretation of their predictive results, *deep soil carbon dynamics is driven more by soil type than by climate*. Although such a result was based on a statistical model with unknown parameters, there was no direct probabilistic judgment to assess the strength of their claim.

Our aim in this article is to scrutinize this claim more closely and check the robustness of the statistical model in view of the many uncertainties: how confident can we be in the effective roles of environmental covariates and climatic factors for the phenomenon under study? What are the respective contributions of signal and noise in what we see? In this paper, we revisit Mathieu's approach under the Bayesian paradigm since Bayesian inference has the advantage of expressing the uncertainties on the unknowns throughout the statistical analysis. We re-parametrize the model to obtain more directly interpretable parameters, change the error term structure to clarify the different sources of uncertainties, and weight the influence of the climatic and environmental drivers for prediction.

A Bayesian selection approach is hereby used in order to quantify the contribution of climatic and environmental factors to soil carbon dynamics. Several Bayesian selection approaches for linear models have been developed in the literature such as: Variable Selection for Regression Models (**VSRM**) (Kuo and Mallick, 1998), Gibbs Variable Selection (**GVS**) (Dellaportas and Ntzoufras, 1997) and Stochastic Search Variable Selection (**SSVS**) (George and McCulloch, 1993).

These methods were applied within the framework of the linear model, where y_i is the outcome response for individual i ($i = 1,\dots,n$) predicted by p potential explanatory covariates x_{ij} for $j = 1,\dots,p$. The intercept is expressed by α and the measurement error by e_i.

$$y_i = \alpha + \sum_{j=1}^{p} \theta_j x_{i,j} + e_i \quad e_i \sim N(0,\sigma^2),$$

with $N(\mu,\sigma^2)$ referring to the Normal distribution with mean μ and variance σ^2. In frequentist selection methods, each variable combination corresponds to a different model, so the variable selection chooses among all possible models the best sub-model based on criteria for model selection such as: AIC, BIC and Mallows's C_p. For a large number of covariates p, it is not computationally achievable to consider all 2^p possible sub-models.

The idea of Bayesian variable selection is to define a binary variable I_j which indicates whether a covariate x_j is influential ($I_j \neq 0$) or not influential ($I_j = 0$) for the response y. I_j is generated from a Bernoulli prior.

The **VSRM** and **GVS** selection methods set $\theta_j = I_j \times \beta_j$. For **VSRM**, I_j and β_j are considered as independent and β_j is sampled from a vague normal prior (Kuo and Mallick, 1998). For **GVS**, β_j is sampled from a conditional prior that depends on I_j such as a Gaussian mixture prior: $P(\beta_j|I_j) = (1-I_j)N(\mu,S^2) + I_j N(0,\tau^2)$, where μ, S^2 and τ^2 are hyperparameters chosen to ensure good mixing of the Monte Carlo Markov Chains (**MCMC**) (Dellaportas and Ntzoufras, 1997). Therefore, these two Bayesian selection methods enable the best sub-model to be selected by affecting null regression coefficients ($I_j = 0 \Rightarrow \theta_j = 0$) for the non influential predictors.

SSVS considers a "slab and spike" prior which depends on I_j for the regression coefficients β_j, with a spike around 0, and a flat slab elsewhere. Then if I_j is null, we assign a value close to 0 for θ_j, which means that the corresponding covariate x_j has no effect on response y. This method was chosen for the present study. The major difference between the scope of the original

SSVS and our specific case is that George and McCulloch (1993) designed the method to select explanatory covariates directly linked to observed data whereas we will specify its use on latent layers. Furthermore, we will evaluate the ability of **SSVS** to handle categorical covariates which are more the rule than the exception when dealing with environmental data.

The paper is organized as follows: Section 2 describes the soil database and model structure, and introduces the Bayesian variable selection to be applied to the latent variables of our non linear multivariate hierarchical model. Section 3 focuses on **SSVS**: first, its performances and limitations are exemplified on three sets of artificial data for a simple linear model with independent quantitative covariates, correlated quantitative covariates and independent mixed covariates. Then, **SSVS** is applied to the entire real data with the complex hierarchical model. Section 4 compares the result of the Bayesian selection model (**SSVS**) to that of a model including all covariates via cross validation. In addition, this section highlights the challenges encountered by applying **SSVS** and suggests how to set up solutions and extensions for this approach. The final section briefly sums up our findings concerning the applicability of **SSVS** in our case study.

2. Materials and Methods

2.1. Data

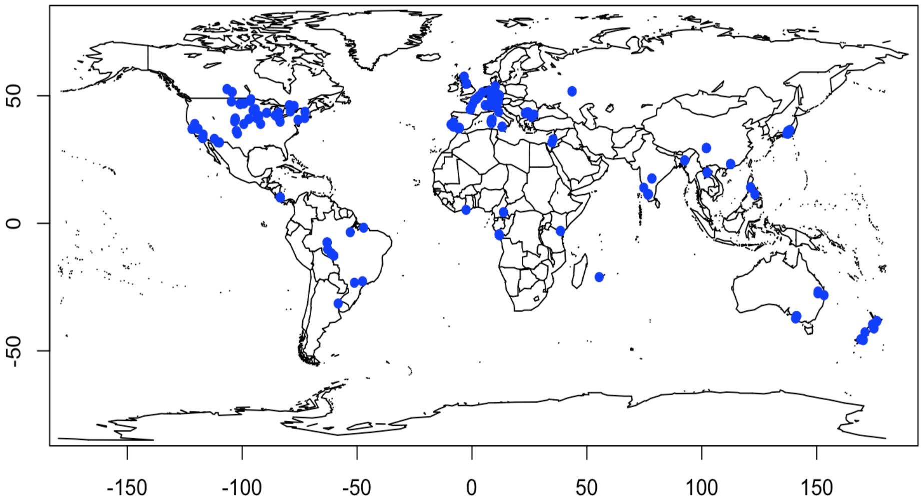

FIGURE 1. *Geographical locations of soil $F^{14}C$ sites.*

Out of the 344 profiles extracted from 87 articles in the soil science and archeology/paleoclimatology literature that constitute a database of available radiocarbon profiles of soil organic carbon (Mathieu et al., 2015), we selected 159 profiles from 50 articles. Several units are used to report radiocarbon concentration. We chose here the $F^{14}C$ unit as recommended by (Reimer, 2004) for environmental samples. $F^{14}C$ is a normalized radiocarbon concentration by reference to the atmospheric radiocarbon content in 1950. For a given site, each record of radiocarbon is given for a soil layer characterized by the depths of its top and bottom levels. Such a preliminary data cleaning was based on the following criteria: i-) the radiocarbon data must have been acquired on bulk organic carbon (not on specific fraction, nor specific molecule), ii-) sites must contain more than 3 observations. Figure 1 shows the site locations where radiocarbon data at various depths were collected. The number of observations varied from one site to another (from 3 to 88 measurements per site). For each of the 159 profiles, the following information of interest is provided: sampling year, location, climate, soil type, land use, organic carbon content and radiocarbon. Soil texture is not considered as it is poorly recorded in many articles from the literature. More details on the database can be found in (Mathieu et al., 2015). In this study, the potential climatic and environmental explanatory covariates are as follows:

- Mean annual precipitation (**MAP**), mean annual temperature (**MAT**), aridity index (**AI**), and absolute shift between July and January temperatures (Δ**T**) are included as representative of the average climate and seasonality of the site. The aridity index, defined by UNEP as the ratio of annual precipitation to annual potential evaporation, was obtained from the FAO 10-minute mean climate grids for global land areas for the period 1950–2000 (Trabacco and Zomer, 2009).
- Latitude (**Lat**).
- The atmospheric radiocarbon of the sampling year (^{14}***C*****atm**).
- Soil type with 13 different categories ordered alphabetically: andosol, arenosol, cambisol, chernozem, ferralsol, fluvisol, gleysol, kastanozem, luvisol, nitisol, phaeozem, podzol, vertisol. We pooled phaeozem and kastanozem soil types into chernozem due to similar characteristics, as they are poorly present in the database. Hereafter, soil type will be considered as a categorical variable with 11 levels.
- Vegetation and land use were combined to form a new factor dubbed "ecosystem", with originally 9 categories distinguished as follows: cultivated-field, cultivated-forest, cultivated-grassland, forest, natural, natural-desert, natural-forest, natural-grassland and natural-savanna. We pooled natural-desert into the "natural" ecosystem. Ecosystem will therefore be considered as a categorical variable with 8 levels.

Among the 159 profiles collected, 55 with missing climatic or environmental covariates were removed from the database. After previous data cleaning, the dataset finally includes 104 sites

TABLE 1. *Contingency table of pairwise combinations of levels between soil type and ecosystem. Abbreviation "C" in column names refers to Cultivated and "N" to Natural.*

	C-Field	C-Forest	C-Grassland	Forest	Natural	N-forest	N-Grassland	N-Savanna	Total
Andosol	0	2	1	0	1	4	0	0	8
Arenosol	0	2	0	0	1	0	0	1	4
Cambisol	2	0	0	1	0	4	2	0	9
Chernozem	2	0	0	0	0	0	11	0	13
Ferralsol	0	0	0	1	0	9	1	2	13
Fluvisol	2	0	0	2	0	0	0	0	4
Gleysol	2	1	0	0	1	0	0	0	4
Luvisol	4	0	2	3	0	7	11	0	27
Nitisol	1	0	0	0	0	4	0	0	5
Podzol	0	3	1	0	0	6	1	0	11
Vertisol	3	0	3	0	0	0	0	0	6
Total	16	8	7	7	3	34	26	3	104

and 951 records. The dataset results from an observational study, which may lead to some confusion due to the spurious association between the correlated and/or the poorly contrasted covariates. The very small number of observations for pairwise- combinations of factors (even a null number for many of them) rules out the possibility of including interactions between soil type and ecosystem in the model (see Table 1). In addition, we anticipate a poor precision of the estimates of the effects of categorical covariates since their design matrix, shown by Table 1, is unfortunately very strongly unbalanced.

To illustrate the composition of the dataset, the boxplots in Figure 2 show the average $F^{14}C$ variation versus the mean levels of non overlapping soil layers, for the most frequent types of profiles collected. This figure only shows average profiles for some specific combinations and prevents any strict interpretation as the number of observations differs from top to depth, and as soil horizon width differs from one profile to another (we do not expect the intensity of processes to be the same at the same depth between two profiles). Figure 2 shows as expected that the radiocarbon decreases with depth: with higher input, topsoil **OM** is more rapidly renewed (and thus shows a younger age) than deep soil **OM**.

2.2. *A multivariate hierarchical non linear model*

The statistical model structure that mimics (eqs 1 and 2) variations of $F^{14}C$ with depth along a profile within a given site is similar to the one considered in Mathieu et al. (2015). It differs only in the homogeneous variance for the measurement error and in the unit chosen to report radiocarbon concentration.

Let $S = 104$ be the total number of carbon soil profiles under study. We note m_s the number of measurements available for site s. Therefore, for each site $s \in \{1 : S\}$ and each depth $x \in \{1 : m_s\}$, the $F^{14}C$ content experimental record $y(s,x)$ is modeled by:

$$y(s,x) = g\left(\phi(s),x\right) + \varepsilon(s,x), \quad \varepsilon(s,x) \sim N(0,\sigma^2) \tag{1}$$

$$g\left(\phi(s),x\right) = \phi_1(s) + (\phi_2(s) - \phi_1(s))exp\left[-\left(\frac{x}{\phi_3(s)}\right)^{\phi_4(s)}\right] \tag{2}$$

As indicated in Fig 3, the structure of the previous statistical model is interpreted:

- ϕ_1 represents F^{14}C in deep soil,
- ϕ_2 refers to the topsoil $F^{14}C$,
- ϕ_3 is related to the depth at half maximum of the $F^{14}C$ peak,
- ϕ_4 describes the more or less rapid decrease of $F^{14}C$.

The ε terms represent the within-site discrepancies between the observed and the adjusted F^{14}C profiles.

To express the variability between the different sites, a linear link is considered between each of

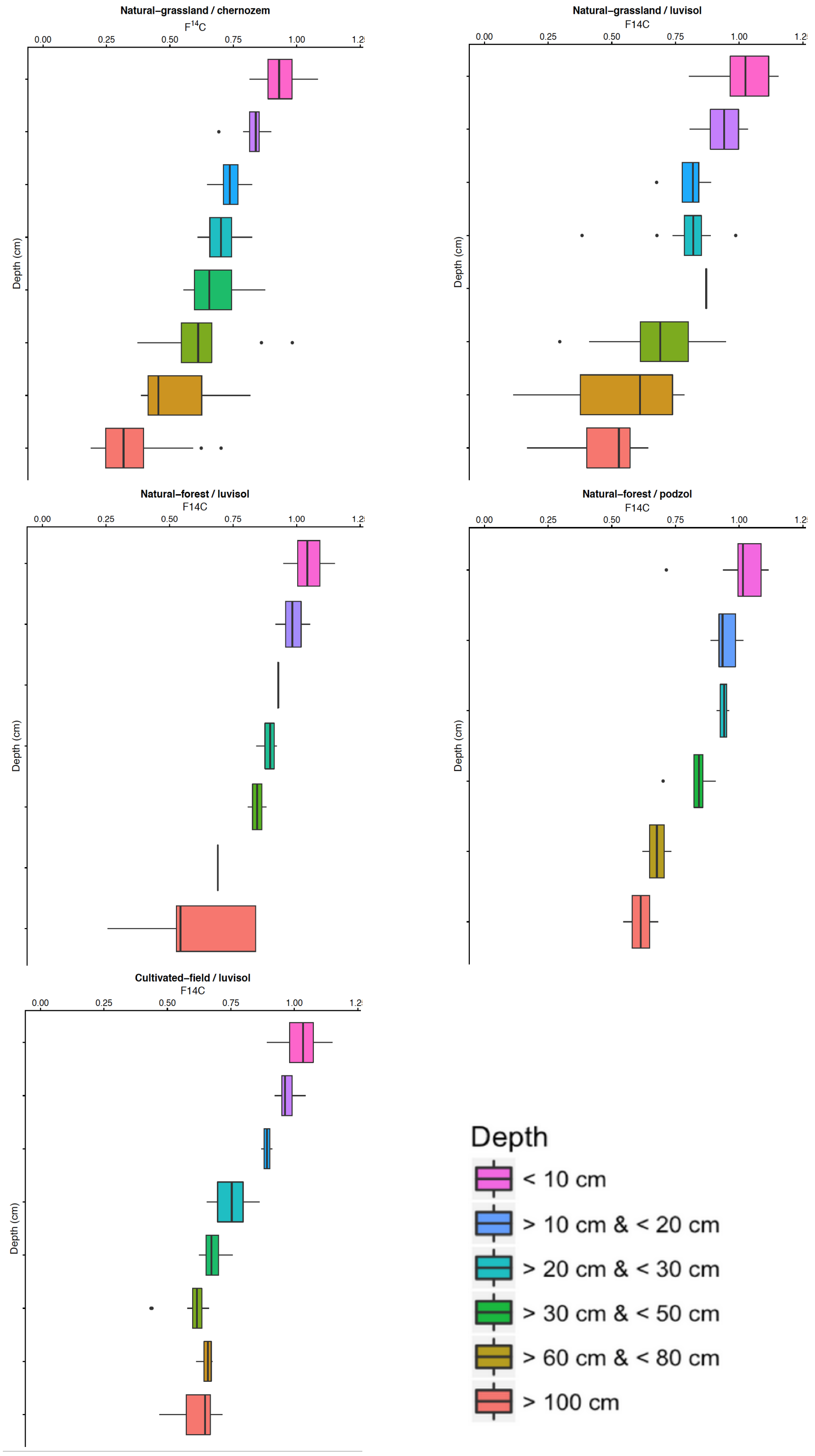

FIGURE 2. *The variation of radiocarbon versus depth is represented on boxplots for the most frequent combinations of ecosystem and soil types. Natural-grassland / chernozem (11 profiles), Natural-grassland / luvisol (11 profiles), Natural-forest / luvisol (9 profiles), Natural-forest / podzol (6 profiles), Cultivated-field / luvisol (4 profiles)*

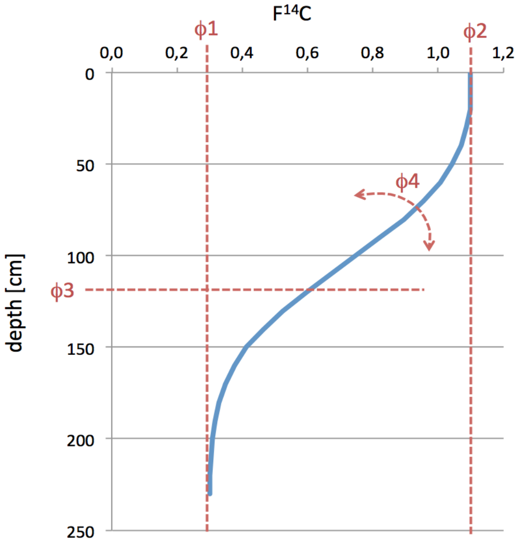

FIGURE 3. *Statistical profile of soil $F^{14}C$ versus depth obtained from Eq. 2.*

the four latent variables $\phi_1(s), \phi_2(s), \phi_3(s), \phi_4(s)$ and the explanatory climatic and environmental variables. We assume that the latent variables are *a priori* independent with a design matrix $X \in M_{P,4}(\mathbb{R})$ defined using a treatment contrast (one level for each categorical covariate is considered as a baseline), as a solution for the redundancy problem due to the presence of categorical variables (soil type and land use) in the linear layer models (without interactions). To be more specific, X is the design matrix with the following form:

$$X = \begin{pmatrix} & & {}^{14}Catm(1) & \text{MAT}(1) & \dots & \Delta T(1) \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1_{soil} & 1_{eco} & {}^{14}Catm(\text{s}) & \text{MAT(s)} & \dots & \Delta T(s) \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ & & {}^{14}Catm(\text{S}) & \text{MAT(S)} & \dots & \Delta T(S) \end{pmatrix}$$

As a first trial, the four latent variables were estimated after a least square optimization to adjust (for each site independently) the curve of Fig 3 to the observations. The estimated variables $\hat{\phi}_1, \hat{\phi}_2, \hat{\phi}_3$ and $\hat{\phi}_4$ were linked to X by four regressions in order to have a preliminary estimation of the regression effects. The diagnostic plots for the linear model led us to perform logarithmic transformations of ϕ_3 and ϕ_4 in order to provide a better agreement with the homogeneous variance hypothesis.

$$\phi_i = X\beta_i + E_i, \quad E_i \sim N_S(0, \sigma_i^2 I) \quad i = 1,2 \tag{3}$$

$$log(\phi_i) = X\beta_i + E_i, \quad E_i \sim N_S(0, \sigma_i^2 I) \quad i = 3,4 \tag{4}$$

$\beta_i = (\beta_{i1}, \dots, \beta_{iP})' \in \mathbb{R}^P$, where $i = 1,2,3,4$, represents the fixed covariate effect relative to each latent variable, and $E_i \in \mathbb{R}^P$ the corresponding centered random effect. ϕ_i and E_i are defined as the following vectors: $\phi_i = (\phi_i(1), \phi_i(2), \dots \phi_i(S))'$ and $E_i = (E_i(1), E_i(2), \dots, E_i(S))'$. In this case study, the number of columns P in the design matrix X is equal to 23 ($P = 1 + (11-1) + (8-1) + 6$). In fact, "1 + (11-1) + (8-1)" is the dimension of the two-way explanatory subspace spanned by the categorical factors "Soil type " and "Ecosystem" that includes the constant. 6 is the number of quantitative regressors. The quantitative regressors in X are normalized to allow comparison of their effects in a rescaled unit. Due to the presence of dummy variables generated by the two categorical factors, the number of columns of the design matrix (23) is greater than the number of explanatory covariates (6+2).

Bayesian selection model: The variable selection procedure is expected to reveal the most influential explanatory variables for the assemblage of the four latent sub-models given with 2 categorical covariates and 6 quantitative ones by equation 3. The idea is to consider a "slab and spike" prior (Dellaportas et al., 2000) for each β_i parameter, with a spike centered at 0, and a flat slab elsewhere. Each combination of included variables corresponds to a different model, so variable selection amounts to choosing among all possible 2^P sub-models if the model considered were a simple linear model with P regressors. For a large number of covariates P , it would be therefore not feasible to consider each possible model separately. In our case, it may seem at first glance that $P = 8$, leading to only $2^8 = 256$ sub-models for each of latent model given by Eqs.3 and 4. Hence the idea of a Bayesian variable selection, where we consider a stochastic exploration of this immense combinatorial set of possible models (O'Hara et al., 2009).
In this article, we concentrate on the **Stochastic Search Variable Selection** introduced by George and McCulloch (1993). This approach is applied to the latent layers ϕ_1, ϕ_2, ϕ_3 and ϕ_4, in presence of categorical covariates.
For the selection procedure, we need to define an indicator variable I_{ij} where i = 1,2,3,4 and j = $1, \dots, P$ as follows:

$$I_{ij} = \begin{cases} 1 & \text{if variable } X_j \text{ has an effect on } \phi_i \\ 0 & \text{otherwise} \end{cases}$$

The mixture prior for β_{ij} depends on I_{ij}:

$$\mathbb{P}(\beta_{ij}|I_{ij}) = (1 - I_{ij})N(0, \tau_{ij}^2) + I_{ij}N(0, c_{ij}^2\tau_{ij}^2) \tag{5}$$

where $i = 1,2,3,4$ and $j = 1, \dots, P$. Based on this Gaussian mixture, τ_{ij} must be small, in order to sample β_{ij} around 0 in situations when variable X_j is not influential, but not strictly restricted to zero, though, otherwise the Gibbs sampler will rarely be able to flip from $I_{ij} = 0$ to visit $I_{ij} = 1$. Furthermore, c_{ij} must be large enough for β_j to be given a flat prior when X_j is needed in the model. A semi-automatic approach to selecting τ_{ij} and c_{ij} was proposed by George and McCulloch (1993) considering the interaction point and relative heights at 0 of the marginal densities. They recommended "good" choices for the couple $(\sigma_{\beta_{ij}}/\tau_{ij}, c_{ij})$, where $\sigma_{\beta_{ij}}$ is the observed standard error associated with the least squares estimate $\hat{\beta_{ij}}$. However, a more appropriate prior for β suggested later is the *hyper-g prior* proposed by Liang et al. (2008) based on the *g-prior* introduced by Zellner (1986). This extension of the *g-prior* has been widely studied and widely used in a regression context. The specification of g is mostly based on a model selection criterion such as the Akaike Information Criterion (AIC, see Burnham et al. (2011)), the Bayesian information criterion (BIC, see Bhat and Kumar, 2010), the Deviance Information Criterion (DIC, see Spiegelhalter et al., 2002), etc. Here, the β prior can be understood as a mixture of spike and slab of *g-priors*. In order to specify g and to ensure a reasonable order of magnitude for β, the hierarchical model without the selection step is first adjusted with a hyper-g prior (with a vague uniform prior at the upper level of the hierarchy). The value of g will be fixed as the posterior mean of this preliminary estimation and used afterwards for the Bayesian selection approach. In that respect, when $I_{i,j}$ is equal to 1, $\beta_{i,j}$ will be generated from the following *g-prior* $N(0, g_i\sigma_i^2(X'X)^{-1}_{j,j})$, to be considered as the slab prior. In contrast, according to the concept of the spike prior, which should be more centered at 0, the $\beta_{i,j}$ corresponding to $I_{i,j} = 0$, will be generated from a *g-prior*, where the variance is much smaller $N(0,(1/c) * g_i\sigma_i^2(X'X)^{-1}_{j,j})$. The hyperparameter c is specified by the user based on a model comparison with different values of c according to the previously cited selection model criteria or to a cross validation study. A hyper prior can also be proposed for c (uniform prior).

The model for Bayesian selection of variables can be finally summed up as follows:

- Likelihood:
  for each site $s \in \{1 : S\}$ and each depth $x \in \{1 : m_s\}$:

$$y(s,x) \sim N(g(\phi(s),x),\sigma^2) \quad with \quad \phi(s) = (\phi_1(s),\phi_2(s),\phi_3(s),\phi_4(s))$$

- Latent variables:

$$\phi_i \sim N_S(X\beta_i,\sigma_i^2 I) \quad i = 1,2$$
$$\log(\phi_i) \sim N_S(X\beta_i,\sigma_i^2 I) \quad i = 3,4$$

  with $\phi_i = (\phi_{1,i},\ldots,\phi_{s,i},\ldots,\phi_{S,i})$, $\phi_i \in \mathbb{R}^P$.

- Priors:

  - $1/\sigma^2 \sim G(0.001,0.001)$
  - $1/\sigma_i^2 \sim G(0.001,0.001)$ for i = 1, 2, 3 and 4
    G(,) refers to the gamma distribution.
  - An intercept is always included and common across all sub-models, for j = 1,2,3,4
    $\beta_{j1} \sim N(0,10000)$
  - for quantitative covariates $j = 2,\ldots,K$
    - $\beta_{1j}|I_{1j} \sim (1-I_{1j}) * N(0,\frac{g_1\sigma_1^2(X'X)^{-1}_{j,j}}{c_1}) + I_{1j} * N(0,g_1\sigma_1^2(X'X)^{-1}_{j,j})$
    - $\beta_{2j}|I_{2j} \sim (1-I_{2j}) * N(0,\frac{g_2\sigma_2^2(X'X)^{-1}_{j,j}}{c_2}) + I_{2j} * N(0,g_2\sigma_2^2(X'X)^{-1}_{j,j})$
    - $\beta_{3j}|I_{3j} \sim (1-I_{3j}) * N(0,\frac{g_3\sigma_3^2(X'X)^{-1}_{j,j}}{c_3}) + I_{3j} * N(0,g_3\sigma_3^2(X'X)^{-1}_{j,j})$
    - $\beta_{4j}|I_{4j} \sim (1-I_{4j}) * N(0,\frac{g_4\sigma_4^2(X'X)^{-1}_{j,j}}{c_4}) + I_{4j} * N(0,g_4\sigma_4^2(X'X)^{-1}_{j,j})$

For $j = 2,\dots,K$ and $i = 1,2,3,4$:

$$I_{ij} \sim \mathscr{B}(p_{ij} = p) \quad \text{with } \mathscr{B}(.) \quad \text{the Bernoulli distribution} \tag{6}$$

i.e. all models are *a priori* equiprobable.

- For the categorical covariates numbered $j = K+1,\dots,P$, with covariate C_j having n_j levels, the algorithm ensures that the n_j modalities are either taken or dropped all together during Monte Carlo Markov Chain (**MCMC**) iteration:

  - for each level $s = 1,\dots,n_j$:
    - $\beta_{1s}|I_{C_j,1} \sim (1-I_{C_j,1}) * N(0,\frac{g_1\sigma_1^2(X'X)^{-1}_{j,j}}{c_1}) + I_{C_j,1} * N(0,g_1\sigma_1^2(X'X)^{-1}_{j,j})$
    - $\beta_{2s}|I_{C_j,2} \sim (1-I_{C_j,2}) * N(0,\frac{g_2\sigma_2^2(X'X)^{-1}_{j,j}}{c_2}) + I_{C_j,2} * N(0,g_2\sigma_2^2(X'X)^{-1}_{j,j})$
    - $\beta_{3s}|I_{C_j,3} \sim (1-I_{C_j,3}) * N(0,\frac{g_3\sigma_3^2(X'X)^{-1}_{j,j}}{c_3}) + I_{C_j,3} * N(0,g_3\sigma_3^2(X'X)^{-1}_{j,j})$
    - $\beta_{4s}|I_{C_j,4} \sim (1-I_{C_j,4}) * N(0,\frac{g_4\sigma_4^2(X'X)^{-1}_{j,j}}{c_4}) + I_{C_j,4} * N(0,g_4\sigma_4^2(X'X)^{-1}_{j,j})$

  For $j = k+1,\dots,P$ and $i = 1,2,3,4$:

$$I_{C_j,i} \sim \mathscr{B}(p_{C_j,i} = p)$$

All levels of a categorical factor therefore receive the same prior selection probability, but more informative priors can be designed, if prior expertise is available to tune the respective importance of the explanatory variables.

The calculation of the posterior distributions of the parameters is based on **MCMC** algorithms such as the Metropolis-Hastings and Gibbs Sampler (Dellaportas et al., 2000). The **SSVS** is easily implemented in **JAGS** (Just Another Gibbs Sampler), as exemplified in Ntzoufras et al. (2002, pp.13-17).

3. Results and Discussion

3.1. Performing SSVS on artificial data

In this section, we illustrate the performance of **SSVS** on latent layers for artificial data generated according to the non linear multivariate statistical structure model (1)+(2)+(5)+(6) when:

1. all independent covariates are quantitative;
2. all covariates are quantitative, and some of them are correlated;
3. the covariates are mixed: some are quantitative and the others are categorical.

The purpose of this artificial data generation is to understand and study the challenges in the application of **SSVS** when the selection aims at hidden sub-models and the model structure is more complex than a simple univariate regression.

SSVS on latent layer models with independent quantitative covariates:

- Example 1: The artificial dataset mimics the real one by taking the same number of sites (104 sites) and depth measurements (951 records). In this example, 6 quantitative (continuous) predictors are considered. The predictors are generated as independent standard normal vectors, $X_1, \dots, X_6$ *iid* $N_{104}(0,1)$, so that they are practically uncorrelated. The regression effects are set to $\beta_1 = (0,1,0,1,0,1), \beta_2 = (0,0,1,1,0,0), \beta_3 = (1,0.8,0,0.7,0,1)$ and $\beta_4 = (1,0,0,1,0.8,0.8)$ with standard deviations $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 0.1$ and $\sigma = 0.1$. The intercept is equal to 1 and will always be kept in the proposals of the latent layer models.

SSVS on latent layers with correlated quantitative covariates:

- As shown in Fig 8, for the real case, covariates may be correlated. Example 2 is designed to illustrate how **SSVS** reacts in the presence of high collinearity. The only difference with example 1 is that the matrix design X contains 2 correlated explanatory variables. X_5 and X_6 are defined as follows:

$$\begin{aligned} X_5 &= 2 \times X_3 \\ X_6 &= X_2 + 1.5 \times Z, \quad Z \sim N(0,1) \end{aligned}$$

SSVS on latent layers with mixed covariates:

- Example 3 introduces categorical variables: this time, the latent linear models ϕ_1, ϕ_2, ϕ_3 and ϕ_4 contain 6 quantitative $(X_1, \dots, X_6)$ covariates and 2 qualitative factors (F_1 and F_2) with respectively 8 and 11 levels. Contrast-sum coding was considered to remain coherent with the presence of quantitative covariates. Regression effects were set to $\beta_1 = (\mathbf{1}, \mathbf{0}, 0, 1, 0, 1, 0, 1)$, $\beta_2 = (\mathbf{0}, \mathbf{1}, 0, 0, 1, 1, 0, 0)$, $\beta_3 = (\mathbf{1}, \mathbf{1}, 1, 0.8, 0, 0.7, 0, 1)$ and $\beta_4 = (\mathbf{0}, \mathbf{0}, 1, 0, 0, 1, 0.8, 0.8)$. **0** and **1** are the index vectors of length 7 or 10 with 0 and 1's corresponding to categorical covariates (position 2 and 3 of the regression coefficients vector). The first position in $\beta_1, \beta_2, \beta_3$ and β_4 is always equal to 1 and refers to the intercept. σ_i, i =1,2,3,4 and σ are fixed as in Example 1. Similar to real data, the experimental design of artificial data is strongly unbalanced.

3.1.1. Sensitivity analysis of the prior for ***SSVS*** *latent layers on independent quantitative covariates*

In order to suggest reasonable values of g_1, g_2, g_3 and g_4 for the spike and slab *g-priors* on the regression effect parameters, the inference of the linear model with a vague uniform prior ($g_i \sim U(10, 1000)$, i = 1,2,3,4) on g was run. The posterior means of g_1, g_2, g_3 and g_4 were plugged into the **SSVS** model.

The prior inclusion probability was fixed to 0.5 in the paper of George and McCulloch (1993). This choice is common for Bayesian selection models since it ensures for all explanatory covariates the same probability of being included in the model. Yet, this prior is informative and favors sub-models with half of the covariates included. For the purpose of studying the impact of the inclusion probability p on the selection results, the **SSVS** was tested under three different prior specifications:

TABLE 2. *The DIC for three proposed priors on probability selection: 1- p is fixed at* 0*, 2- a Beta prior on p* $\mathbb{B}(2,2)$ *3- a uniform prior on p* $U(0,1)$ *. Models with smaller DIC should be preferred to models with larger DIC.*

p	DIC
fixed to 0.5	-1511
beta prior	-1515
uniform prior	-1512

TABLE 3. *The posterior inclusion probability for the most frequent models among the 3000 MCMC iterations for* ϕ_1, ϕ_2, ϕ_3 *and* ϕ_4 *latent linear models. The three proposed priors succeed in identifying correctly the best sub-models.*

Most frequent model	False detection	$p = 0.5$	beta prior	uniform prior
X_2, X_4, X_6	0	0.73	0.66	0.63
X_3, X_4	0	0.44	0.46	0.48
X_1, X_2, X_4, X_6	0	0.78	0.61	0.52
X_1, X_4, X_5, X_6	0	0.72	0.53	0.44

1. p is fixed to 0.5 for all covariates,
2. a Beta distribution prior on p ($p \sim \mathbb{B}(2,2)$),
3. a uniform distribution prior on p ($p \sim U(0,1)$).

For these three tested models, the $c_i (i = 1,2,3,4)$ were fixed to 100 for the four latent linear models. According to the Deviance Information Criterion (DIC) easily provided by JAGS, the **SSVS** with a beta prior on p is preferred (see Table 2)

For a linear model with a large number of covariates, a uniform inclusion probability of 0.5 may bias the best sub-model by being too complex since it favors the sub-models with half of the covariates selected. Figure 4 gives the total number of selected covariates identified among MCMC iterations for the third latent linear model that involves 6 covariates. This result highlights that the choice of 0.5 promotes the selection of sub-models with 3 covariates. The Beta and Uniform distributions prior increase the probability selection of sub-models with more than half the number of total covariates.

According to the result obtained, a prior Beta distribution will be proposed on the inclusion probability p for the further **SSVS** models.

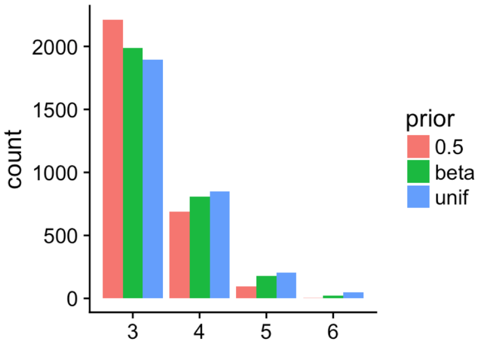

FIGURE 4. *The number of selected covariates identified among the MCMC iterations (nb of iterations = 3000) for the third latent variables* (ϕ_3) *for the three proposed priors for the probability selection p.*

TABLE 4. ***SSVS** evaluation for artificial data including only independent quantitative covariates. Panels a, b, c, d are the results obtained for ϕ_1, ϕ_2, ϕ_3 and ϕ_4 latent layers, respectively. Rows give the three most visited sub-models. Columns correspond to the different tested priors. F.C. (False Choice) sums up both false inclusion and false exclusion. Prob. is the probability appearance of model subsets throughout iterations. The best sub-models detected by the **SSVS** with the three proposed values of c do not contain any false detection.*

	c = 10		c = 100		c = 1000		c = 5000	
	Prob	F.C	Prob	F.C	Prob	F.C	Prob	F.C
a)	0.31	-	0.66	-	0.81	-	0.84	-
	0.13	X_3	0.09	X_3	0.06	X_1	0.06	X_3
	0.12	X_5	0.08	X_1	0.05	X_3	0.04	X_5

	c = 10		c = 100		c = 1000		c = 5000	
	Prob	F.C	Prob	F.C	Prob	F.C	Prob	F.C
b)	0.30	-	0.46	-	0.40	-	0.39	-
	0.10	X_2	0.17	X_2	0.25	X_2	0.28	X_2
	0.08	X_1	0.06	X_6	0.08	X_6	0.07	X_6

	c = 10		c = 100		c = 1000		c = 5000	
	Prob	F.C	Prob	F.C	Prob	F.C	Prob	F.C
c)	0.33	-	0.61	-	0.72	-	0.74	-
	0.23	X_3, X_5	0.16	X_5	0.13	X_5	0.12	X_5
	0.21	X_3	0.14	X_3	0.10	X_3	0.11	X_3

	c = 10		c = 100		c = 1000		c = 5000	
	Prob	F.C	Prob	F.C	Prob	F.C	Prob	F.C
d)	0.29	-	0.53	-	0.58	-	0.59	-
	0.25	X_2, X_3	0.18	X_2	0.17	X_2	0.18	X_2
	0.21	X_2	0.17	X_3	0.16	X_3	0.15	X_3

3.1.2. *Sensitivity analysis prior for SSVS latent layers on independent quantitative covariates*

In this section, we test the "best" choice of the hyperparameter c for the β prior specification. We consider the following values of c: 10, 100, 1000 and 5000. The **MCMC** is run for $30,000$ iterations after a burn-in of $10,000$ iterations. In addition, a Beta prior $\mathbb{B}(2,2)$ is proposed for the inclusion probability p. The four panels in Table 4 show, for Example 1 of artificial data, the **SSVS** performance under different priors on $\beta_1, \beta_2, \beta_3$ and β_4. These tables show the three most frequent models with the false inclusion (False positive) or exclusion (False negative) rates of predictors.

For the different spike and slab priors, **SSVS** performs extremely well for c_i = 10, 100 and 1000 ($i = 1,2,3,4$) since the best sub-models identified for each of the four latent layers contain

TABLE 5. *Comparison between the three **SSVS** models with different values of c according to the DIC criterion. The best model is identified by the lowest DIC estimation.*

c	DIC
10	-1513
100	-1515
1000	-1523
5000	-1520

no false detections (see the first line of the panels a), b), c) and d)). The best sub-models do not contain any false choice. As expected, as the value of c increases, the posterior distribution becomes more peaked, which can be explained by the increase in probability appearance along these settings. In fact, the probability of the most visited model increases with higher values of c (see the probability values in the first row of the previous four tables). For example, in Table 2–(d), the best sub-model under $c = 10$ is visited 870 times throughout $30,000$ iterations, while the best sub-model under $c = 5000$ is visited 1770 times. The **SSVS** with $c = 1000$ is identified as the best according to the DIC estimations. Moreover, a vague uniform prior can be proposed on parameter c in order to have a better estimation. Generally speaking, **SSVS** performs well on latent layer models with independent quantitative covariates.

3.1.3. The presence of collinearity increases false detection on ***SSVS*** *in the latent layer*

George and McCulloch (1993) showed that collinearity may reduce the efficiency of **SSVS** by increasing the number of promising models in a linear model framework. Collinearity between some covariates in a latent layer model can also increase the rate of false positives/negatives especially when one of the correlated covariates is influential but the other is not. The **SSVS** model is now considered with a Beta prior on the probability selection p $(p \sim \mathbb{B}(2,2))$ and a vague uniform prior on c $(\mathbb{U}(5,1000))$.

Figure 5 illustrates how correlated covariates restrict **SSVS** performances. The **SSVS** model provides a probability judgment about the most frequent explanatory covariates combination. In addition to that, the **SSVS** also provides a probability judgment about the inclusion of each of the explanatory covariates on the different sub-models identified throughout MCMC iterations. Here, the Posterior Inclusion Probabilities (PIP) for each covariate separately are illustrated in Fig.5. In the first and third panels, the selected covariates correctly specify the influential covariates taken *a priori* into account to generate artificial data. Outputting, both X_3 and X_5 as non influential, and X_2 and X_6 as influential for ϕ_1 was expected since the correlated covariates were *a priori* both influential/not influential at the same time. With regard to the second panel, ϕ_2 was generated taking into account X_3, while X_5 is omitted *a priori*. Therefore as X_5 is correlated with X_3, **SSVS** misleads and selects X_5. Likewise, X_2 and X_3 were not taken into account when generating ϕ_4. As a result, two false choices are reported, the exclusion of X_5 and the inclusion of X_2.

3.1.4. ***SSVS*** *performance within latent layer mixed covariates (quantitative and qualitative)*

The algorithm for mixed covariates was developed to give the same inclusion probability to all levels of the same categorical covariate. The results obtained in Example 3 highlight some limitations of **SSVS** with regards to the presence of categorical covariates in the latent layer. It can be clearly seen that **SSVS** may fail to detect some influential explanatory categorical covariates. However, **SSVS** does not seem to induce false choice inclusion. In our case study, it considers a categorical covariate as influential only if it is actually influential: it can miss some of them but does not induce false positives.

The new dummy covariates needed to handle the presence of categorical covariates F_1 (8 levels) and F_2 (11 levels) strongly increase the dimensions of the space of competing models to be

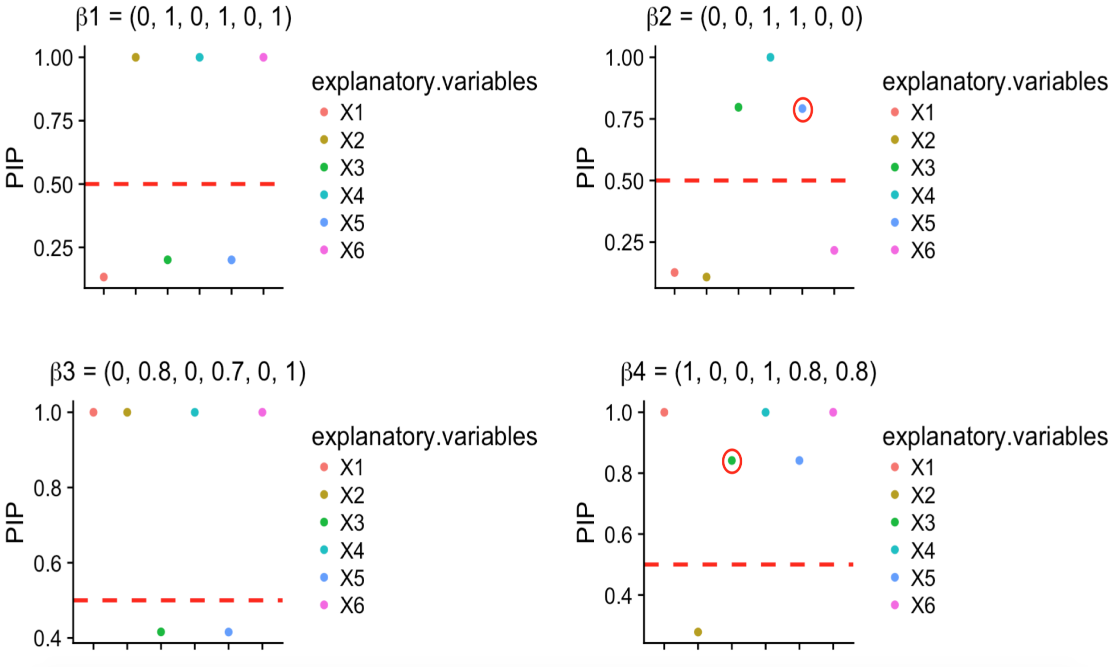

FIGURE 5. ***SSVS*** *evaluation for artificial data including both independent and correlated quantitative covariates. Panels a, b, c, d give the results obtained for the Posterior Inclusion Probability (PIP) separately for each covariate throughout the sub-models identified by the MCMC iterations for the four latent layers. A PIP higher than 0.5 indicates a strong probability of inclusion of the relative covariate in the best sub-model. The title of each graph reflects the true value of the regression coefficients from which artificial data were generated: e.g. in Example 2,* X_5 *was correlated with* X_3*, and* X_6 *with* X_2*. The red circle reflects false detection, for example false inclusion for* X_5 *and* X_2 *respectively in the* ϕ_2 *and* ϕ_4 *latent models.*

stochastically explored. The selection results summarized in Table 6 were obtained after applying the **SSVS** algorithm on the artificial data generated as Example 3:

TABLE 6. *The selection results obtained by applying the* ***SSVS*** *on latent layers with mixed explanatory covariates. For each latent layer, the real model from which the data was generated, the best sub-model detected with the highest frequency of appearance throughout MCMC iterations and the false negative detections are given.*

latent layers	real model	best model detected by **SSVS**	false negative	probability appearance
ϕ_1	F_1,F_2,X_2,X_4,X_5	F_1,F_2,X_2,X_4,X_5	0	0.765
ϕ_2	F_1,F_2,X_1,X_2,X_3	F_1,F_2,X_1,X_2,X_3	0	0.223
ϕ_3	F_2,X_1,X_2,X_5,X_6	X_1,X_2,X_5,X_6	F_2	0.882
ϕ_4	F_2,X_4,X_5,X_6	X_4,X_5,X_6	F_2	0.695

Results displayed in Table 6 show that **SSVS** is able to identify the influential quantitative covariates $(X_1,\ldots,X_6)$ (0 false detection for quantitative covariates). Moreover, for the first and the second latent layers ϕ_1 and ϕ_2, the best sub-models detected by the **SSVS** are correct with null false detections. In contrast, for ϕ_3 and ϕ_4, the categorical covariates F_1 and F_2 are detected as false negative detections respectively for ϕ_3 and ϕ_4 linear models. These results highlight a limitation of **SSVS** related to the presence of categorical covariates in latent layers. It is clear that **SSVS** fails to detect some influential explanatory categorical covariates. However, **SSVS** does not induce false choice inclusion in this case study. In other words, it considers a categorical covariate as influential only if it is actually influential.

Such avoidance of false choice inclusion might stem from the fact that **SSVS** with even prior weights tends to dampen the selection probability of a categorical covariate with a big number of modalities. In fact, the prior distribution of $\beta_k \in \mathbb{R}^M$ when covariate k is selected (i.e. $I_k = 1$) is proportional to $\frac{1}{(g\sigma_k(X_k'X_k)^{-1})^M}$. Consequently, when M becomes large, the prior distribution $P(\beta_k|I_k = 1)$ will vanish to 0. For that reason, **SSVS** may seem to be reluctant to select a categorical covariate with a high number of levels.

3.1.5. Variance sensitivity analysis for SSVS

As mentioned above, George and McCulloch (1993) designed and applied **SSVS** to detect explanatory covariates directly linked to the observed response whereas we applied it to covariates buried in latent layers in the framework of a hierarchical Bayesian model. To complete the assessment in our specific case, we evaluated the sensitivity of **SSVS** to the variance within the latent layer.

Overall, sensitivity variance analyses highlight that an increase in variability between sites (expressed by the $\sigma_1, \sigma_2, \sigma_3$ and σ_4 of the latent layer models) does decrease **SSVS** robustness to select the best subset of covariates.

In our specific case, two sources of variability are to be distinguished: variability between sites expressed by $\sigma_1, \sigma_2, \sigma_3$ and σ_4 and variability within the same site expressed by σ. In order to test **SSVS** sensitivity to intersite variability changes, we simplified the proposed statistical model by fixing ϕ_2, ϕ_3 and ϕ_4. **SSVS** was applied only on ϕ_1, which has a linear effect on the $F^{14}C$ response. We tested **SSVS** for four different values of $\sigma_1 = (0.01, 0.1, 2.5, 3)$. Figure 6 shows the posterior inclusion probability for one of the considered covariates "X_2", for different σ_1 settings. Figure 6 clearly illustrates the impact of σ_1 on the posterior inclusion probability (**PIP**): the more σ_1 increases, the more PIP decreases. It even reaches a PIP close to 0.5 for $\sigma_1 = 3$, leading to a potential false choice (exclusion) of an important variable.

3.2. SSVS on observed radiocarbon profiles

*3.2.1. Application of **SSVS** on soil $F^{14}C$ profiles*

The aim of this section is to highlight the contribution of **SSVS** to understanding which climatic and environmental factors are likely to control soil carbon dynamics. Based on the results obtained on artificial data, it can be claimed that the presence of categorical covariates in the model can produce false exclusions of some of the influential categorical covariates. In addition, the correlation between some covariates such as temperature and latitude, may yield false detection, especially if they do not have the same effect on latent layers as we showed in subsection 3.1.3.

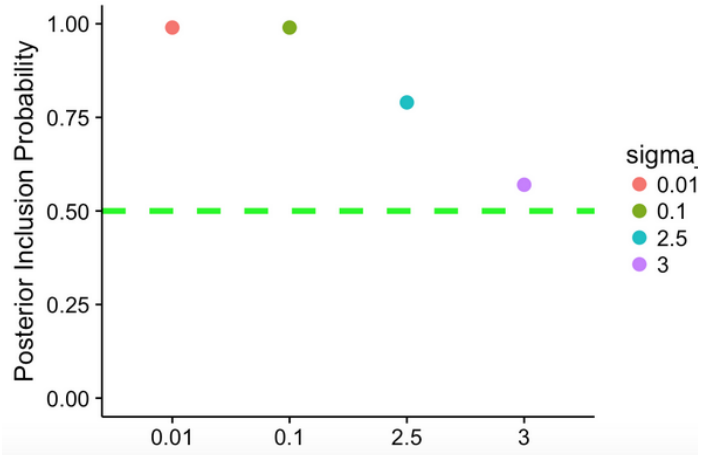

FIGURE 6. *Posterior inclusion probability in relation with σ_1 value. The illustration here is for X_2 included as a descriptor of the ϕ_1 latent layer model, ϕ_i for $i = 2,3,4$ being fixed. The posterior inclusion probability decreases versus increasing values of σ_1. The green dashed line represents the decision-making rule: for a posterior inclusion probability higher than 0.5, the relative predictor is considered as influential.*

Choice of c for regression effects prior

The **SSVS** was applied on real data by considering a beta prior $\mathbb{B}(2,2)$ on inclusion probability with different values of c: 10, 100, 1000 and 5000. Furthermore, the model with uniform prior $\mathbb{U}(10,10,000)$ on $c_i(i=1,\dots,4)$ was also tested.

TABLE 7. *The DIC comparisons for five* **SSVS** *models under different prior specifications for c_i (i =1, 2, 3, 4). The table also summarizes the DIC for the full model containing all explanatory covariates.*

SSVS models	DIC
$c_i = 10$	-1869
$c_i = 100$	-1806
$c_i = 1000$	-1890
$c_i = 5000$	-1855
Uniform prior on c_i	-1860
Full model	-1726

All **SSVS** models returned a better adjustment than the full model, according to the **DIC** criterion. The best model is identified by the lowest value of **DIC**. The SSVS model on radiocarbon profiles will thus be established with c_i equal to 1000 for i = 1,. . . ,4.
To investigate the predictive power of the **SSVS** models, a cross validation procedure was conducted. **SSVS** models were adjusted on the same learning sets (70% of studied sites) and 30% of data were used as validation sets. The average Posterior Relative Errors (PRE) for all sites under the different depth measurements are summarized in the following Table 8. Here, the difference of the PRE among the **SSVS** models is very small. According to the results on Tables 7 and 8, hyperparameter c is to be fixed to 1000.

TABLE 8. *Posterior Relative Error (PRE) computed for all sites and for all depth measurements throughout MCMC iterations. The PRE difference between the models is very small. The best model has the lowest PRE on the validation sets.*

SSVS models	Posterior Relative Error on learning sets	Posterior Relative Error on validation sets
$c_i = 10$	0.225	0.406
$c_i = 100$	0.230	0.402
$c_i = 1000$	0.234	0.413
$c_i = 5000$	0.238	0.416
uniform prior on c_i	0.235	0.417

Results of Posterior Inclusion Probability (PIP) for covariates among the sub-models identified by MCMC simulations

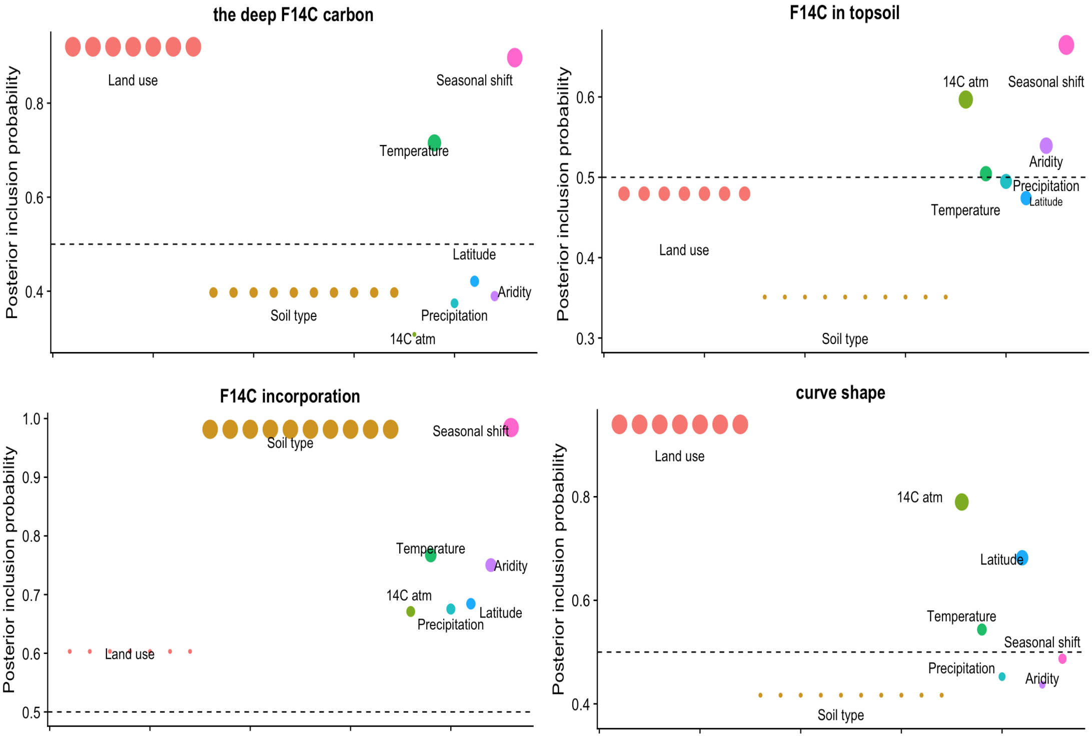

FIGURE 7. *Posterior inclusion probabilities for all explanatory covariates obtained by applying the* **SSVS** *to the entire real database. The size of points depends on the importance of the posterior inclusion probability.*

Panels 1, 2, 3 and 4 of Fig 7, show the Posterior Inclusion Probabilities (PIP) for each categorical covariate throughout the different sub-models visited by the Markov chains. According to the selection results obtained on artificial data with mixed covariates in subsection 3.1.4, the **SSVS** provides a good performance on quantitative covariates (no false detection). However, it can miss some significant categorical covariates. Panels 1 and 2 indicate that the seasonal shift and the temperature are included with probabilities 90% and 73% respectively throughout the visited sub-models for ϕ_1 and ϕ_2. ϕ_1 and ϕ_2 are respectively related to the deep and topsoil $F^{14}C$. All explanatory covariates are selected for ϕ_3 in its latent model. All the categorical covariates (land use or soil type) selected with a probability higher than 0.5 are included in the best sub-model. So, land use is very surely included in the best sub-models of ϕ_1, ϕ_3, ϕ_4 and soil type in the ϕ_3 best sub-models. In contrast, every categorical covariate not selected (PIP smaller than 0.5), may be significant for the model since the **SSVS** approach can yield negative false detection for categorical covariates. For example, soil type is a priori not included in the best sub-model of ϕ_1 but might still be significant to explain deep soil radiocarbon.

Moreover, a posterior probabilistic beliefs on the association of explanatory covariates is provided by looking at the most frequent covariate combinations throughout the **MCMC** iterations (see Table 9).

Results of 2 most frequent combinations of covariates identified by Stochastic Search Variable Selection

TABLE 9. *High 2 frequency models (Model1 and Model2) for each of the latent linear models. It represents the 2 most frequent combinations of explanatory covariates among all the MCMC iterations. The linear models with all explanatory covariates are identified for ϕ_2, ϕ_3 and ϕ_4.*

Latent linear model	High frequency model	frequency (n.iter = 180,000)
ϕ_1	Model1: land use, temperature and seasonal shift	12,549
	Model2: land use, seasonal shift	10,822
ϕ_2	Model1: all covariates	6,606
	Model2: seasonal shift	4,272
ϕ_3	Model1: all covariates	36,819
	Model2: all covariates except land use	12,587
ϕ_4	Model1: all covariates	14,782
	Model2: land use, $F^{14}C$ atmospheric, latitude	7,336

According to the Table 9, the frequency visits to the best sub-models are very small with respect to the total number of iterations (180,000) and maybe not all the sub-models are explored by the MCMC. Moreover, the full models are detected as the best sub-models for three of the latent layers ϕ_2, ϕ_3 and ϕ_4. However, the covariates Posterior Inclusion Probabilities (PIP) highlight that the best model chosen should contain the covariates with a PIP higher than 0.5. Furthermore, for more detailed investigations, the unknown parameters of the statistical radiocarbon model are re-estimated, taking into account all the covariates for which the PIP is higher than 0.5 (see Fig. 7). In addition, as the SSVS may miss the inclusion of some influential categorical covariates, one may wonder whether the soil type has really no effect on the ϕ_1 latent linear linear model or whether it is perhaps simply not detected by the **SSVS** model. The answer to this question is reported in the following table.

Comparison of DIC for 5 sub-models taking into account for some sub-models the drawback of SSVS when categorical covariates are present in the model

TABLE 10. *Model* contains the explanatory covariates with a PIP higher than 0.5. To investigate whether a non selected categorical covariate is significant, we add respectively to Model*, the non included categorical covariates (land use or ,soil type) identified with a PIP smaller than 0.5. The Table displays the DIC criteria comparisons between the different models.*

Models	DIC
Most frequent model (denoted Model1 for each of latent layers in Table 9)	-1703
Model* = the model adjusted on the covariates where their PIP are higher than 0.5 (see Fig.7)	-1837
Model* + considering the soil type for ϕ_1	-1897
Model* + considering the soil type for ϕ_1 and ϕ_2	-1890
Model* + considering the land use for ϕ_2 and soil type for ϕ_1	-1968
Model* + considering the soil type for ϕ_1 and land use for ϕ_2 and ϕ_4	-1879

The **DIC** comparison in Table 10, shows that the best model is the one that includes both PIP> 0.5 detected explanatory covariates, i.e. "soil type" for deep soil radiocarbon (ϕ_1) and "land use" for topsoil radiocarbon (ϕ_2) (DIC = -1968). In addition, this result highlights that the **SSVS** is misleading in that it detects two significant categorical covariates (2 false negatives). The final selection of covariates for the radiocarbon model is summed up in Table 11.

Selection results for the best sub-model: the climatic and environmental factors that affect soil radiocarbon dynamics

TABLE 11. *The final selected covariates for each of the four latent layer models ϕ_1, ϕ_2, ϕ_3 and ϕ_4. For the third latent layer ϕ_3 all explanatory covariates are selected. Furthermore, for ϕ_1 and ϕ_4 four covariates are identified among 8 as significant while 5 covariates are detected for ϕ_2 as influential towards the 8 potential climatic and environmental factors.*

Best model	final selected covariates
ϕ_1	land use, soil type, temperature, seasonal shift
ϕ_2	land use, atmospheric $F^{14}C$, temperature, seasonal shift and aridity
ϕ_3	land use, soil type, atmospheric $F^{14}C$, temperature, aridity, precipitation, latitude and seasonal shift
ϕ_4	land use, latitude, atmospheric $F^{14}C$, temperature

A further point is the correlation among covariates. For example, temperature and seasonal shift are positively correlated (see Fig.8). This could suggest that temperature may not be really influential for ϕ_1 as its inclusion may be the result of its correlation with the highly influential covariate "seasonal shift". However, if we take a look at the second panel of Fig 7, we can see that seasonal shift has an effect on ϕ_2, which is not the case for temperature, indicating that the correlation between temperature and seasonal shift does not seem to affect **SSVS** performance that much.

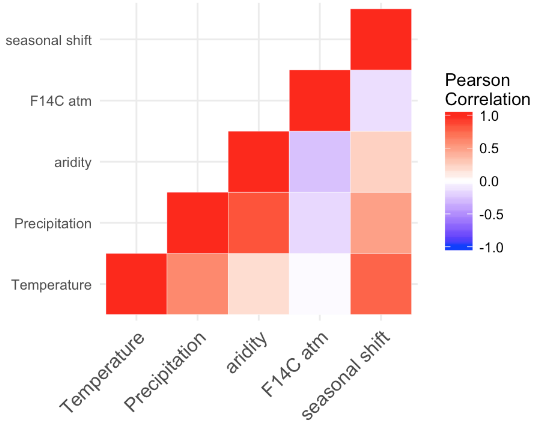

FIGURE 8. *Correlation matrix of the six quantitative explanatory covariates. The darker the color (red or blue), the stronger the correlation between the variables (positive or negative)*

Posterior Predictive Checking

To build additional confidence in our selected model, a predictive posterior check is useful. It compares data replications y^{rep} according to the **SSVS** model (c = 1000 and $p \sim Beta(2,2)$) governed by parameter θ, with the observed data y. The behavior of a model with regard to a feature of interest is quantified by a discrepancy measure $T(y,\theta)$. Here, the $T(y,\theta)$ quantity is the average of the squared difference between y and the non linear predicted mean $g(\theta)$, where $\theta = (\beta_1,\beta_2,\beta_3,\beta_4,\sigma_1,\sigma_2,\sigma_3,\sigma_4)$. After computing T(y,$\theta$) and T($y^{rep}$, θ), a posterior predictive p-value is defined as $Pr[T(y^{rep},\theta) \geqslant T(y,\theta)|y]$ (Gelman et al., 2013). The posterior predictive p-value is not as strictly used as in the classic procedure comparing a statistic with some Type 1 error. Gelman et al. (2013) interpret the posterior predictive p-value as the proportion of data replications according to the proposed model $T(y,\theta)$ that exceeds $T(y^{rep},\theta)$. A model is rejected if the Bayesian p-value is rather small. In our case, the posterior predictive p-value is equal to 0.47! (see Fig.9)

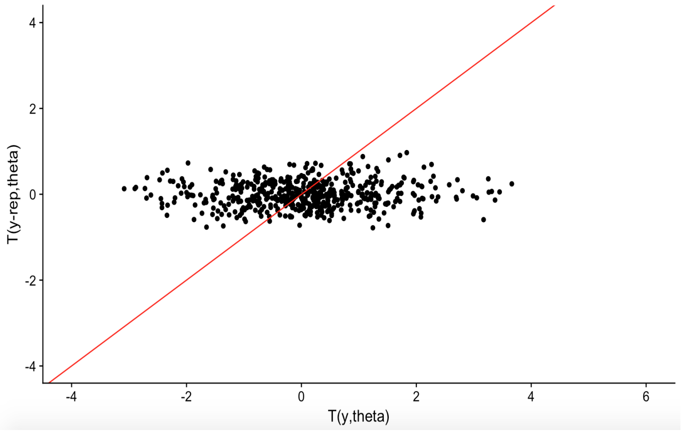

FIGURE 9. *The discrepancy measures $T(X,y^{rep},\theta)$ calculated on replicated data and parameters model θ versus $T(X,y,\theta)$ calculate on real data and θ. The estimated Bayesian p-value is equal to 0.47.*

Better understanding of the climatic and environmental factors that affect soil radiocarbon dynamics

Besides detecting whether a covariate has an influence or not on ϕ_1, ϕ_2, ϕ_3 and ϕ_4 , quantifying the effect of each influential covariate is also of interest. For example, it would be useful to know what happens to ϕ_1 (representing radiocarbon content in deep soil) if there is a strong rise in temperature due to global warming. The answer to this question is given by the posterior distribution of regression coefficients $\beta_1, \beta_2, \beta_3$ and β_4 corresponding to the significant explanatory covariates (see Fig.10 and 11).

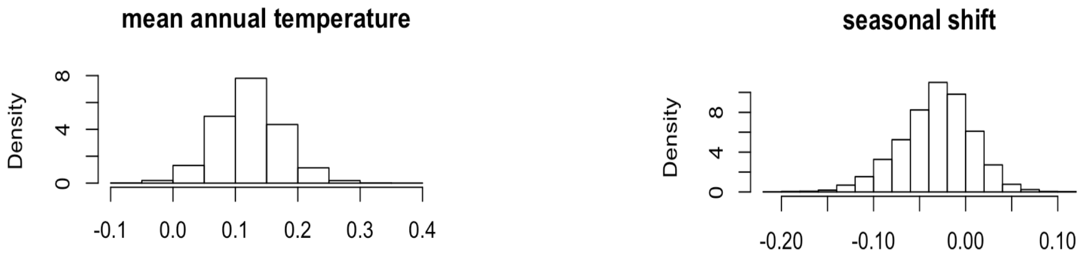

FIGURE 10. *The posterior distribution of the regression effects corresponding to the significant numerical covariates for the deep soil radiocarbon (ϕ_1) latent model: mean annual temperature and seasonal shift.*

TABLE 12. *The significant explanatory numerical covariates for deep radiocarbon with their posterior mean estimations and their posterior probabilities of the sign of their relative effects throughout MCMC iterations.*

Covariates	posterior probability (to be + or -)	posterior mean estimation
mean annual temperature	0.99 (+)	+0.12
seasonal shift	0.80 (-)	-0.03

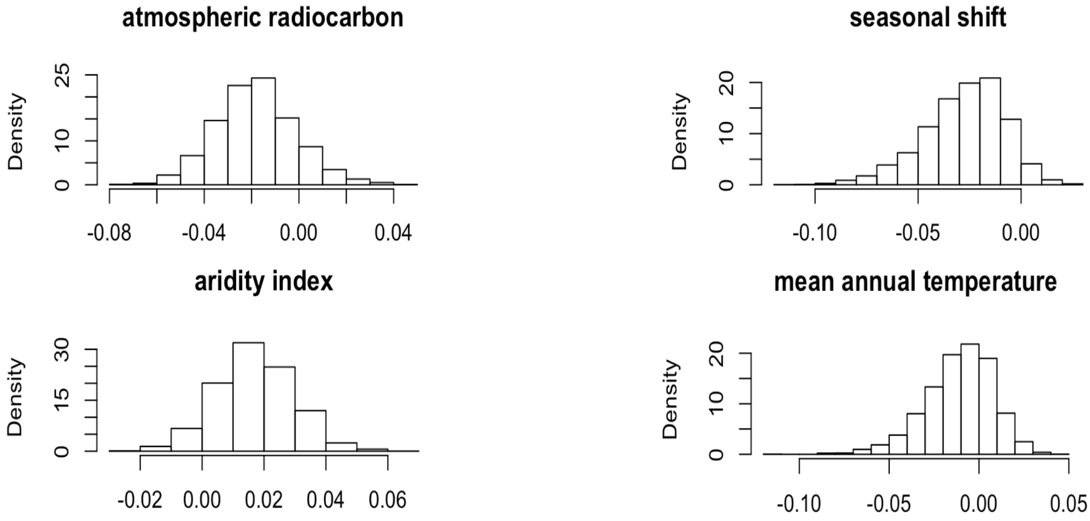

FIGURE 11. *The posterior distribution of the regression effects corresponding to the significant numerical covariates for topsoil radiocarbon latent model (ϕ_2): atmospheric radiocarbon, seasonal shift, aridity index and mean annual temperature.*

TABLE 13. *The significant explanatory numerical covariates for topsoil radiocarbon with their posterior mean estimations and their posterior probabilities of the sign of their relative effects throughout MCMC iterations.*

Covariates	posterior probability (to be + or -)	posterior mean estimation
atmospheric radiocarbon	0.86 (-)	-0.018
seasonal shift	0.95 (-)	-0.028
mean annual temperature	0.70 (-)	-0.011
aridity index	0.92 (+)	+0.017

Interpreting the posterior effect of radiocarbon profiles is not straightforward because of the very high variability of atmospheric radiocarbon concentration with time. A massive change occurred in the 1960s with atmospheric tests of nuclear weapons that doubled the radiocarbon concentration in the atmosphere, leading to a so-called "radiocarbon bomb peak" (see panel a of Fig. 12). Topsoil already incorporates peak-bomb-derived radiocarbon whereas deep soil is still free of radiocarbon enriched components (see panel b of Fig.12). The interpretation of radiocarbon changes differs greatly, therefore, depending on whether it is related to top soil or to deep

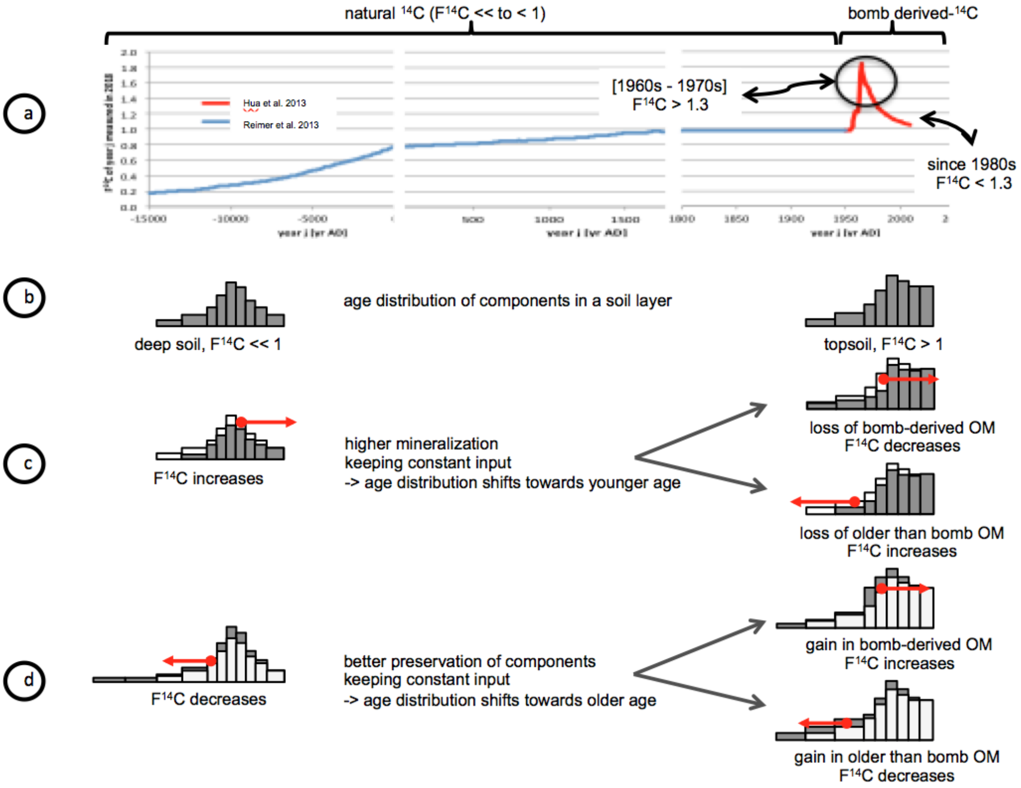

FIGURE 12. *The first graph a shows the variation of the atmospheric $F^{14}C$ concentration over time. The soil was affected specially by above-ground nuclear testing from about 1950 until 1963. Panel b highlights the variation of radiocarbon amount between deep and topsoil. The last 2 panels c and d show the impact of physical processes on deep and topsoil radiocarbon. Furthermore, all panels provide an indication on age distribution since the radiocarbon is an indicator of the mean residence time of soil carbon.*

soil.

An increase in microbial activity that leads to higher mineralization will result in a weaker weight of older components relative to newly input ones in the age distribution of the mixture of soil components within the same soil layer (panel c in Fig.12). This will result in an increase of radiocarbon in deep soil but a decrease in topsoil radiocarbon where the weight of the peak-bomb derived components decreases due to a higher mineralization (panel c in Fig.12). We face the opposite effect in the case of processes that will enhance the organic matter stabilization and will better preserve old material (panel d in Fig.12).

Keeping in mind that point, our results for the deep soil highlight a positive posterior effect of mean annual temperature and a negative posterior effect of seasonal shift. In practice, an increase of 1°C in the mean annual temperature will result in an increase of radiocarbon of 0.12 and an increase of 1°C between the highest and the lowest monthly temperature will result in a decrease of radiocarbon by 0.03. This increase of deep soil radiocarbon with temperature is in agreement with a higher mineralization associated to an enhancement of microbial activity under higher temperature. Likewise the decrease of radiocarbon with seasonality matches what is known about the impact of seasonality on soil dynamics with much younger soils, i.e. with a higher turnover under the tropics than in boreal, i.e. continental areas, where soil shows a much lower turnover and thus yields much lower radiocarbon.

Topsoil is negatively impacted by atmospheric radiocarbon, seasonal shift and mean annual temperature and positively impacted by aridity. Most of the profiles included in the database were sampled posteriorly to the 1960s, i.e. for years during the bomb peak decrease with an overrepresentation of the 1990s. The bomb peak gradually penetrates into soil layers with a time lag that depends of the mean residence time of components in the different layers. With a mean residence time of 100 yrs, the maximum of $F^{14}C$ will be in the early 2000's. Thus, the negative impact of the atmospheric F^{14}C reflects the fact that an increase in the atmospheric F^{14}C means that sampling was made some years before, when the bomb peak had not yet reached its maximum in soil. The dilution effect of bomb-peak derived components is thus higher, yielding a lower (closer to 1) mean radiocarbon. However, this effect remains very low (-0.01 decrease of topsoil radiocarbon associated to an increase of atmospheric radiocarbon by 1) reflecting the dilution effect of the bomb-peak and the disequilibrium of the database in which sites sampled in the 1990s are overrepresented. Negative impacts of seasonal shift and mean annual temperature by -0.02 and -0.01 respectively are the counterpart for topsoil of what is observed for deep soil. An higher mineralization for the mean annual temperature, leading for an higher loss of bomb-derived organic matter and a better preservation for seasonal shift yielding for a relative gain of the oldest components. It is noteworthy that impacts for topsoil appear much smaller than for deep soil. This result is counter-intuitive and no reason for that can be advanced. The positive impact of aridity is in agreement with a well-known low microbial efficiency in arid environments compared to humid ones. An increase in aridity results in a better preservation of the bomb-peak derived components and thus to an increase in the topsoil radiocarbon. The effect of aridity remains very low at +0.01.

A large difference exists between the magnitudes of the posterior estimation of the influential covariates of the latent variable for topsoil and deep soil. While an explanation stemming from the database disequilibrium can be put forward to explain the low magnitude of atmospheric radiocarbon, no clear evidence can be provided for the other covariates.

4. Extensions and challenges

Database: To better predict the evolution of soil carbon dynamics with climate change and land use change practices, there is a need to collect more data for the type of soil (arenosol, fluvisol and gleysol) and ecosystem (natural/savanna, cultivated/grassland and forest) about which we do not have much information. In this study, the experimental design was strongly unbalanced, which affects the precision when estimating the quantities of interest: ϕ_1, ϕ_2, ϕ_3 and ϕ_4. Furthermore, optimization of the experimental design should take into account the type of contrast used to solve the redundancy of the model caused by the presence of categorical explanatory covariates. An interesting new track will be to know where to take new samples and for which climatic and environmental conditions in order to improve the overall estimation. Another issue associated with data is correlation. Some of the explanatory covariates are naturally correlated (see Figure 8).

For example, the aridity index (**AI**) is proportional to the mean annual precipitation (**MAP**) by definition (see eq. 4) since:

$$AI = \frac{MAP}{ET_p} \qquad ET_p : \text{potential evapotranspiration rates}$$

SSVS is sensitive to the presence of correlated covariates as already seen in Section 3.1.3 (see Fig 8). More investigation can be done considering other Bayesian predictive criteria for model selection according to the paper by Piironen and Vehtari (2017).

Improving the Bayesian selection model. The test carried out on artificial data shows that **SSVS** does not always detect influential categorical explanatory covariates. This issue could be thoroughly explored using the Bayesian effect fusion approach introduced by Pauger and Wagner (2017). They proposed a Bayesian approach for a sparse representation of the effect of a categorical predictor in linear models. The originality of their work is that it not only allows selection of categorical covariates but also induces fusion among the categorical covariate levels which have essentially the same effect on the response. Besides this approach, Bayesian variable selection for group Lasso presented in the paper by Xu et al. (2015) selects variables both at the group level and also within a group. Revisiting the traditional Bayesian approach to the group Lasso problem, they developed a Bayesian group Lasso model with spike and slab priors for problems that also require selection of categorical explanatory variables.

5. Conclusion

In this paper, we have discussed the performance and limitations of **SSVS** on latent layers in the framework of a hierarchical Bayesian model applied to soil radiocarbon. The results on artificial data show that collinearity may lead to false inclusion or exclusion in the best sub-model selected. Besides collinearity, if variability on the latent model response is high, the posterior inclusion probability may blur the effect of influential explanatory covariates as exemplified in Section 3.1.5. Furthermore, **SSVS** is not always able to select the influential categorical covariates, but at least does not seem to consider a covariate as influential unless it is indeed the case. Despite the complexity of **SSVS** compared to the full model, we show that the Bayesian selection approach has a better adjustment and prediction level in our case study. Finally, the application of **SSVS** to soil $F^{14}C$ profiles highlighted the influence of soil types on soil carbon dynamics by impacting deep soil $F^{14}C$, topsoil $F^{14}C$ and $F^{14}C$ incorporation. Our results also indicate that temperature affects deep soil $F^{14}C$ more than topsoil.

6. Acknowledgments

We are indebted to two anonymous reviewers for their attentive and detailed reviews and for their numerous constructive comments. The data set was kindly provided by Christine Hatté. This work is supported by the research contract ANR 14-CE-01-0004 and is part of the first author's PhD study financed by CEA.

References

Batjes, N. H. (1996). Total carbon and nitrogen in the soils of the world. *European Journal of Soil Science*, 47(2):151–163.

Bhat, H. S. and Kumar, N. (2010). On the derivation of the bayesian information criterion. *School of Natural Sciences, University of California*.

Burnham, K. P., Anderson, D. R., and Huyvaert, K. P. (2011). Aic model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, 65(1):23–35.

Carvalhais, N., Forkel, M., Khomik, M., Bellarby, J., Jung, M., Migliavacca, M., Mu, M., Saatchi, S., Santoro, M., Thurner, M., et al. (2014). Global covariation of carbon turnover times with climate in terrestrial ecosystems. *Nature*, 514(7521):213–217.

Dellaportas, P., J. and Ntzoufras (1997). On Bayesian model and variable selection using MCMC. *Technical report, Departement of Statistics, Athens University of Economics and Business*.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2000). Bayesian variable selection using the Gibbs sampler. *Biostatistics-Basel-*, 5:273–286.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

He, Y., Trumbore, S. E., Torn, M. S., Harden, J. W., Vaughn, L. J., Allison, S. D., and Randerson, J. T. (2016). Radiocarbon constraints imply reduced carbon uptake by soils during the 21st century. *Science*, 353(6306):1419–1424.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Martin, M., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., and Arrouays, D. (2011). Spatial distribution of soil organic carbon stocks in france. *Biogeosciences*.

Mathieu, J. A., Hatté, C., Balesdent, J., and Parent, É. (2015). Deep soil carbon dynamics are driven more by soil type than by climate: a worldwide meta-analysis of radiocarbon profiles. *Global Change Biology*, 21(11):4278–4292.

Ntzoufras, I. et al. (2002). Gibbs variable selection using BUGS. *Journal of statistical software*, 7(7):1–19.

O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117.

Pauger, D. and Wagner, H. (2017). Bayesian effect fusion for categorical predictors. *Preprint arXiv:1703.10245*.

Piironen, J. and Vehtari, A. (2017). Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735.

Reimer, P. J. (2004). Intcal04. *Radiocarbon*, 46(3):1029–1058.

Scharpenseel, H. (1971). Radiocarbon dating of soils–problems, troubles, hopes. *Paleopedology: Origin, Nature and Dating of Paleosols. papers*.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Stocker, T. (2014). *Climate change 2013: the physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Trabacco and Zomer (2009). Global aridity index (global-aridity) and global potential evapo-transpiration (global-pet) geospatial database. *CGIAR Consortium for Spatial Information*.

Xu, X., Ghosh, M., et al. (2015). Bayesian variable selection and estimation for group Lasso. *Bayesian Analysis*, 10(4):909–936.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*.