Vertical Distribution of carbon in Soils - Bayesian Analysis of carbon content and C14 profiles

Rana Jreich

université PARIS-SACLAY

NNT : 2018SACLV060

Thèse de doctorat

Distribution verticale du carbone dans les sols - Analyse bayésienne des profils des teneurs en carbone et de ^{14}C

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université de Versailles-Saint-Quentin-en-Yvelines

Ecole doctorale n°129 Sciences de l'environnement d'Ile-de-France
Spécialité de doctorat : Météorologie, océanographie, physique de l'environnement

Thèse présentée et soutenue à Paris, le 28 novembre 2018, par

RANA JREICH

Composition du Jury :

M. Martyn Plummer Professeur, University of Warwick	Rapporteur
M. Joël Chadoeuf Directeur de Recherche, INRA (Avignon)	Rapporteur
M. Philippe Bousquet Professeur, Université Versailles Saint-Quentin en Yvelines	Président
Mme Delphine Derrien Maître de conférences, INRA (Nancy)	Examinatrice
Mme Christine Hatté Chercheuse-ingénieure E6, CEA-LSCE	Directrice de thèse
M. Éric Parent Ingénieur Général des Ponts, des Eaux et des forêts (AgroParisTech)	Co-Directeur de thèse
M. Jérôme Balesdent Directeur de Recherche, INRA (Aix-en-Provence)	Invité

First, I would like to express my sincere gratitude to my advisors, Christine HATTÉ and Eric Parent. The first day of our meeting, in the "train bleu" restaurant at Lyon station, stayed etched in my memory. The atmosphere was very relaxed and confortable. In this day, i did not know that the Lyon station will be the departure station to take the train to start my research and have a final destination to defense my thesis successfully. I thank Eric for sharing with me his expertise in Bayesian inference and for all the scientific discussions addressed. These discussions allowed me to gain more experience and perspective in statistics. On the other hand, I mostly appreciated the multidisciplinary aspect of my thesis. I was so lucky to have Christine Hatté as my thesis supervisor. Thank you for helping me discover the new field of soil carbon with finest pedagogy. You surprised me by your reflection on statistics methods which remarkably helped improving soil carbon modeling. Your positive vibes gave me confidence and intensified my motivation. Sharing many moments outside our usual work environment made me able to discover both Eric and Christine's wonderful personalities and their loving hearts.

I also thank the members of my thesis jury, Prof. Philippe Bousquet, Prof. Martyn Plummer, Research Director Joël Chadeouf, Dr. Delphine Derrien and Research Director Jérôme Balesdent, for devoting their time to read the manuscript and for their insightful comments and remarks.

My thesis was the opportunity to create friendly relationships with Annarosa, Anna Bilkova, Martina, Marie, Rapahelle, Sema, Félix, Mathieu, Thimothé, Saint-Clair and Yann from AgroParisTech. Not to forget my former colleagues Nadia Ben Abdallah, Ha trung and Sandra Negro, with whom I shared a lot of good times. My warm gratitude goes to Laurine, Claire, Kristan, Laurie, Naoufal, Marion, Hugo, Marion P., Marion L. , Marine, Virginie, Romain, Alison, Annouck, Mirian Valente, etc... from LSCE. I would also like to thank my other colleagues at MIA-AgroParisTech and LSCE for their constructive and pleasant working relationship.

A heartfelt thanks to my loyal friend and my soul mate Jihane Aouni. You are the reason that I am here in France. From the beginning, you encouraged me to follow my dream and told me that nothing is impossible in life. We came to France for continuing our studies and we went forward step by step, always together. You are the one who brings out the best in me.

There is no success in life without Family. I want to thank the lord God for having such a beautiful and warm family. I dedicate my thesis to all the member of my family, my father Tony, my mother Najah, my three brothers Henri, Robert, Joe, my lovely sister Mireille , my brother in Low Paul Kiame, and to the soul of my grandmother Wadad (may her soul rest in peace). You all have filled me with your love and positives vibes. A special thanks for my angels Jospeh and Rita, for coming from Belgium to support your aunt, I love you more than you could ever imagine.

CONTENTS

List of Figures

List of Tables

ACRONYMS

A_{SN}	Standard activity, background-corrected and δ13 C-normalized (the 1950 atmospheric activity)
A_{ON}	sample activity, background-corrected and δ13 C-normalized
AI	Aridity Index
BEF	Bayesian Effect Fusion
BGL-SS	Bayesian Group Lasso with Spike and Slab
BSGS	Bayesian Sparse Group Selection
BUGS	Bayesian inference Using Gibbs Sampling
$^{\circ}$C	Celsius degree
C	Carbon
^{13}C	Carbon-13
^{14}C	Carbon-14
$\Delta^{14}C$	Radiocarbon fractional deviation from the standard activity
cal. yr BP	Calibrated years before present
CO	Carbon monoxide
CO_2	Carbon dioxide
CRU	Climatic Research Unit
Dif_P	Seasonal precipitation shift
Dif_T	Seasonal temperature shift
ESM	Earth System Model
FAO	Food and Agriculture Organization
$F^{14}C$	Ratio of the sample activity to the standard activity measured in the same year, both activities background-corrected and δ13 C-normalized
$F^{14}C$ atm	Atmospheric $F^{14}C$ of the sampling year
g/m^2	Gram per square meter
HBM	Hierarchical Bayesian Modeling
HWSD	Harmonized World Soil Database
IPCC	Intergovernmental Panel on Climate Change
JAGS	Just Another Gibbs Sampler
MAP	Mean Annual Precipitation
MAT	Mean Annual Temperature
max_P	maximum precipitation
max_T	maximum temperature
min_P	minimum precipitation
min_T	minimum temperature
^{14}N	Nitrogen-14
NPP	Net Primary Production
PgC	Petagram of Carbon
pMC	Percent Modern Carbon
SSVS	Stochastic Search Variable Selection
UNEP	United Nations Environment Programme
WRB	World Reference Base
%wt	Percent Weight
yr BP	Year Before Present

Le réchauffement climatique est une menace pour tous les écosystèmes terrestres et océaniques et pour l'adaptation de l'homme à son milieu. Selon la dernière évaluation du Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC), la température moyenne de la surface de la terre a augmenté d'environ 0.9 °C entre 1901 et 2012 (Stocker, 2014). Le réchauffement climatique est principalement dû aux émissions de gaz à effet de serre, spécialement le dioxyde de carbone, le méthane et le protoxyde d'azote. La concentration du dioxyde de carbone a augmenté de 45% depuis la période pré-industrielle en conséquence des activités humaines qui déséquilibrent le cycle global du carbone (Harris, 2010).

Le sol est un réservoir majeur de carbone et joue un rôle essentiel dans le système climatique.

Le cycle du carbone est largement décrit : sa bonne compréhension constitue la contribution première du travail du GIEC. Une description est fournie par Ciais et al. (2014) (Figure 1).
En résumé, l'océan reste le plus grand réservoir de carbone avec environ 900 PgC pour l'océan en surface et 39 000 PgC pour l'océan profond, l'atmosphère contient actuellement 828 PgC selon Prather et al. (2012) (environ 590 PgC en période pré-industrielle). La végétation piège entre 450 et 650 PgC (Prentice et al., 2001) et le sol contient entre 1500 et 2400 PgC sous forme de matières organiques(Batjes, 1996).

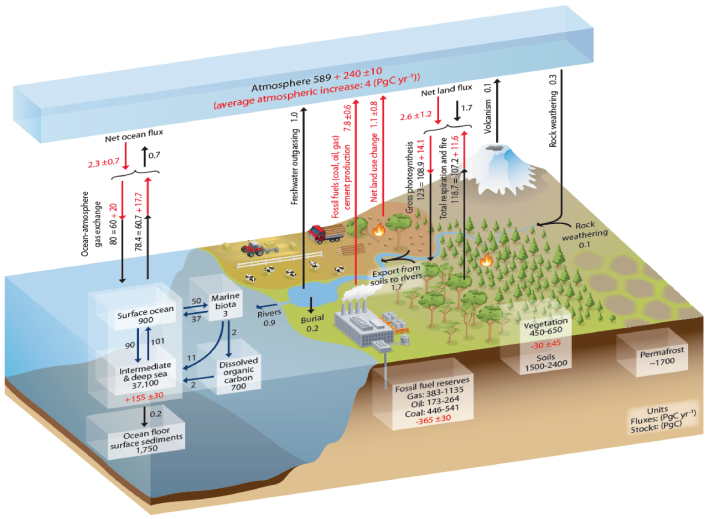


Figure 1: Le cycle global du carbone. Les flèches rouges représentent les flux anthropiques et celles représentent les flux naturels (Ciais et al., 2014).

En outre, Batjes (1996) souligne le manque de connaissances en ce qui concerne le carbone du sol profond, vu que peu d'échantillons ont été prélevés dans les niveaux profonds du sol. Par ailleurs, Batjes (1996) met en évidence que la difficulté de l'estimation globale du réservoir de carbone du sol provient d'une part de la confusion des effets du climat, de la végétation et de l'usage de sol et d'autre part de lacunes dans les données pour représenter tous les types de sols et les conditions climatiques et environnementales. De plus, une étude récente de Tifafi et al. (2018), basée sur trois base de données mondiales, montre une grande différence dans les estimations régionales et globales des stocks de carbone dans le sol (Figure 2). Le stock total du carbone de sol est estimé à environ 3 400 PgC par SoilGrids, alors qu'il est d'environ 2 500 PgC selon la base de données harmonisée des sols du monde (HWSD). Toutefois, le carbone du sol est un réservoir beaucoup plus grand que l'atmosphère et, avec la végétation, constitue le seul réservoir sur lequel l'homme peut intervenir afin d'augmenter sa capacité de stocker plus de carbone, et ainsi piéger une partie du CO_2 dérivé des combustibles fossiles émis.

Les sols contribuent le plus aux échanges de carbone avec l'atmosphère.

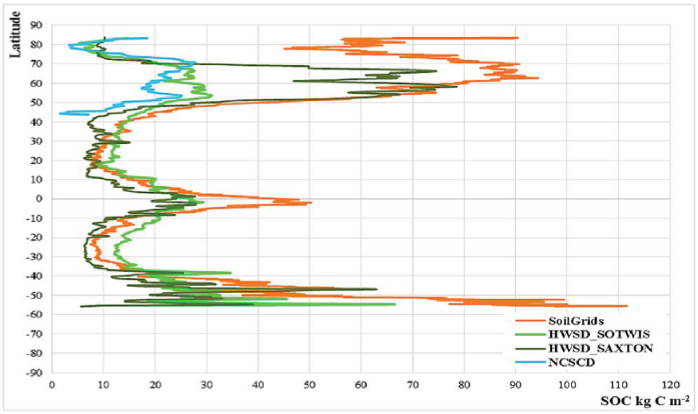


Figure 2: Stock de carbone total (kg C m^{-2}) sur la couche supérieure (0,1 m) par latitude et pour les différentes bases de données (Tifafi et al., 2018).

Le cycle du carbone est un système dynamique échangeant des flux de carbone d'un réservoir à l'autre. La perturbation anthropique du cycle global du carbone entraîne une augmentation des émissions de carbone vers l'atmosphère. Les responsables principaux en sont les combustibles, la production de ciment (émission de 7.8 $\pm$ 0.6 PgC par an) et le changement d'usage des sols (émission de 1.1 $\pm$ 0.8 PgC par an). La végétation piège approximativement 108.9 PgC par an de carbone atmosphérique via la photosynthèse, dont 14.1 PgC sont d'origine anthropique. Une partie de ce carbone est intégrée dans le sol via les racines des plantes et les feuilles mortes tandis que l'autre partie est renvoyée dans l'atmosphère via la respiration des plantes (107.2 PgC par an). L'échange net entre biosphère et atmosphère (60 PgC par an, dont 49.4 PgC par an entre atmosphère et le sol) reste le plus important dans le cycle du carbone global.

Les sols ont un rôle dans la lutte contre le réchauffement climatique

Le rapport du GIEC 2013 souligne le rôle clé des sols pour l'atténuation et l'adaptation aux changements climatiques : les échanges nets en carbone les plus importants se produisent entre les sols et l'atmosphère. En outre, l'impact des changements d'usage de terre a été souligné dans une étude réalisée par Deng et al. (2016). Cette méta-analyse basée sur 103 publications récentes de 160 sites dans 29 pays différents a montré que le stock de carbone dans le sol augmente significativement après la conversion de terres agricoles en prairies (0.30 Mg par hectare et par an) et de forêts en prairies (0.68 Mg par hectare et par an), mais diminue de manière significative après la conversion de prairies en terres agricoles (0.89 Mg par hectare et par an) et des forêts en terres agricoles (1.74 Mg par hectare et par an).
Le réchauffement climatique peut être atténué grâce à la gestion des terres cultivées, la gestion des pâturages et la restauration des sols organiques. Des stratégies de pratiques agricoles et de conservation des forêts stimulent l'augmentation de l'absorption du carbone par les sols et permettent de mieux assurer la sécurité alimentaire en préservant la fertilité du sol. À cet égard,une meilleure compréhension du carbone de sol est à l'ordre du jour suite au rapport du GIEC 2013 où, pour la première fois, les sols sont considérés comme l'un des leviers possibles d'atténuation des changements climatiques. Dans ce contexte, nous pouvons souligner l'initiative «4 pour mille» https://www.4p1000.org qui vise à augmenter la séquestration mondiale du carbone, pour les premiers 40 cm du sol, d'un pourcentage de 0.4% par an, afin d'atténuer les problèmes mondiaux liés au changement climatique, à l'insécurité alimentaire et à la pollution de l'environnement (Lal, 2016). Pour distinguer entre séquestration et stockage, il est communément établi que la séquestration doit être durable (au moins 100 ans, comme recommandé par le protocole de Kyoto), alors que le stockage peut être à court terme ou à long terme.

La représentation actuelle de la dynamique du carbone du sol par les modèles mécanistes n'est pas entièrement satisfaisante

Plusieurs modèles mécanistes ont été proposés pour exprimer la variation du carbone du sol en fonction de la profondeur. Cependant, un effort pour mieux représenter les simulations de la dynamique du carbone du sol est nécessaire, d'autant plus que la vision des processus physiques est incomplète. Les modèles de carbone du sol les plus utilisés sont:

1. le modèle de RothC (Coleman et al., 1997) qui intègre également un modèle simple eau/sol,
2. le modèle CENTURY (Parton et al., 1987) qui inclut des modules plus complexes pour la croissance des plantes, la gestion des opérations, etc.

Le modèle RothC simule que les 30 premiers centimètres du sol et le modèle CENTURY simule les premiers 20 centimètres (Falloon and Smith, 2010). Le calibrage du modèle mécaniste pour les premiers centimètres montre la nécessité d'intégrer le carbone du sol profond, vu que le carbone stable et résistant des horizons profonds interagit avec les niveaux de surface et contribue alors au cycle mondial du carbone et aux émissions de CO_2 par suite des changements globaux du climat et des pratiques d'usage de sol. Les différences entre les modèles de carbone du sol incluent dans les modèles du système de la terre sont principalement dues aux différences entre la production primaire net (PPN) et la paramétrisation des sous-modèles de composition de la matière organique du sol. La limitation de la modélisation mécaniste de la dynamique du carbone du sol provient aussi du fait que ces modèles sont paramétrés sous des conditions climatiques et environnementales spécifiques. Cet aperçu sur les modèles mécanistes de la dynamique du carbone du sol souligne l'importance de prendre en compte le carbone total des sols et d'étendre les conceptualisations des processus à toutes les échelles de temps et d'espace. De plus, il existe de grandes incertitudes quant aux processus qui ralentissent la minéralisation et protègent la matière organique du sol. Parmi ces processus, on peut distinguer: l'inaccessibilité spatiale contre les micro-organismes et les enzymes, l'hydrophobicité, l'encapsulation dans les macromolécules organiques, la récalcitrance de la litière, les interactions entre la matière organique et les minéraux, etc. Jusqu'à présent, le principal défi reste d'exprimer ces nouveaux concepts de stabilisation / déstabilisation par des équations différentielles afin de les incorporer dans la modélisation mécaniste. En outre, la majorité des modèles mécanistes du carbone du sol sous-estime la quantité du carbone du sol puisque le carbone profond n'est pas pris en compte dans les bilans du carbone (Houghton, 1995).

Les isotopes du carbone permettent de valider la représentation de la dynamique du carbone dans le sol

La meilleure façon d'évaluer la performance des modèles mécanistes de la dynamique du carbone du sol est de les comparer avec les données empiriques. Ainsi, une comparaison directe entre les résultats du terrain, du laboratoire, des données et des sorties du modèle mécaniste peut être établie. Pour représenter la matière organique, la spécifier, la suivre et donner une cinétique aux processus, des mesures de la matière organique à plusieurs profondeurs sont nécessaires. Tout d'abord, la quantité du carbone dans le sol peut être définie par les données de teneur en carbone produites par le laboratoire analysant des échantillons prélevés du terrain. Ensuite, il existe des méthodes de traçage isotopique comme les traceurs ^{13}C et ^{14}C pour quantifier le temps de résidence de la matière organique du sol, allant de quelques jours jusqu'à plusieurs milliers d'années. La première technique de traçage est fondée sur la surveillance d'abondance du ^{13}C en cas de changement de végétation (des plantes de type C3 en C4 ou vice versa). Malheureusement, les données disponibles à partir de cette technique ne sont pas en nombre suffisant pour l'évaluation du modèle parce qu'un changement de type de photosynthèse de la végétation est exigé. La deuxième technique, la datation au radiocarbone, est plus puissante. Effectivement, le sol est un témoin des variations des concentrations du radiocarbone de l'atmosphère, en particulier la variation due aux essais nucléaires

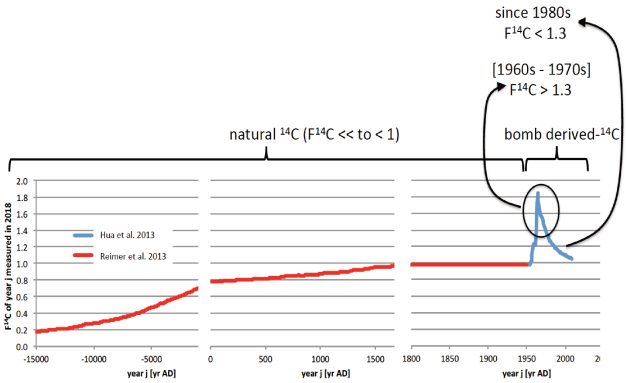


Figure 3: La quantité du ^{14}C de l'année x mesurée en 2018, pour tout matériau atmosphérique. Le sol reproduit la variation du radiocarbone atmosphérique dans le cadre du réservoir de radiocarbone mais, en mélangeant des matières organiques de plusieurs périodes, le signal atmosphérique est fortement atténué dans le sol.

atmosphériques (Figure 3). Le contenu du radiocarbone ^{14}C de l'atmosphère a doublé en 1962 par rapport à son niveau naturel, égal au contenu du radiocarbone de l'atmosphère dans l'hémisphère nord en 1950.

La modélisation statistique est une alternative pour mieux représenter les incertitudes sur la dynamique du carbone des sols

En fait, l'intégration des nouveaux concepts de protection de la matière organique dans les modèles mécanistes, la prise en compte de ce qui se passe dans les couches profondes du sol et l'extension de la conceptualisation des processus à toutes les échelles de temps et d'espace, nécessitent un travail de développement intensif. L'écriture de cette formalisation mathématique détaillée des processus physiques restant encore largement hors de portée aujourd'hui, notre proposition est de construire un modèle statistique qui décrive la dynamique du carbone dans la sol en se fondant sur des données empiriques du radiocarbone et des teneurs en carbone échantillonnées à différentes profondeurs pour des sites répartis sur tout le globe terrestre.

Le modèle statistique a l'avantage d'échapper à la difficulté d'exprimer tous les processus mécanistes qui contrôlent la dynamique du carbone dans le sol. Il permet également de représenter les incertitudes. On propose ici une méta-analyse statistique pour mieux comprendre la dynamique du radiocarbone ainsi que celle des teneurs du carbone. Cette méta-analyse est basée sur 343 sites collectés à partir de 86 articles publiés dans la littérature de pédologie/archéologie et paléoclimatologie. Le pergélisol et l'histosol ont été omis, notre étude porte donc seulement sur les sols minéraux. De plus, pour chaque site, on dispose des informations climatiques (température, précipitation, etc.) géographiques (longitude, latitude, altitude) et environnementales (type de sol et type d'écosystème). La base de données renferme 17 différents types de sol et 9 types d'écosystème (végétation + usage du sol (naturel versus cultivé)). Afin d'améliorer le plan expérimental, certains types de sol partageant des même propriété physiques ont été fusionnés ensemble. L'analyse statistique est mené finalement sur 9 types de sol et 6 types d'écosystème.

Le modèle statistique que l'on propose ici est inspiré du modèle statistique décrit par Mathieu et al. (2015) afin de représenter les variations du radiocarbone en fonction de la profondeur. Ce modèle statistique est caractérisé par une structure hiérarchique non linéaire à effets aléatoires. Afin d'intégrer les informations climatiques et environnementales, les effets aléatoires on été reliés par des liens linéaires aux facteurs qui peuvent potentiellement impacter la matière organique du sol. Dans cette étude, on considère 9 variables climatiques et environnementales parmi 33 variables possibles. Par exemple, les températures enregistrées pour tous les 12 mois d'années seront

résumées par la température annuelle moyenne et l'écart saisonnier de la température entre le mois le plus chaud et le plus froid. Les 9 variables explicatives ont été choisies de manière à réduire le problème de multicolinéarité que l'on peut rencontrer dans ces cas. En fait, quelque soit l'ensemble explicatif retenu, on ne peut pas échapper à l'association naturelle qui existe entre climat, type de végétation et type de sol.

Le premier défi de ma thèse est d'apporter des réponses aux questions des scientifiques du sol et de leur rendre la vision de la dynamique du carbone dans le sol plus claire. En fait, notre analyse statistique servi, dans un premier temps, à proposer une structure de modèle qui réalise un bon ajustement des données empiriques et qui nous permette de générer un profil du carbone pour des endroits du globe où aucune donnée n'a été observée. Les scientifiques du sol cherchaient à identifier les facteurs climatiques et environnementaux dont les effets contrôlent la dynamique du carbone dans le sol et à distinguer si ces facteurs impactent le carbone en surface et en profondeur de la même façon. Pour répondre à cette question, plusieurs méthodes bayésiennes de sélection de variables numériques et catégorielles ont été explorées afin de quantifier numériquement et d'une manière probabiliste les effets des facteurs climatiques et environnementaux.

Description de la base de données

La base de données utilisée pour notre méta-analyse est détaillée dans le premier chapitre du manuscrit. A l'origine, les données de 343 sites ont été collectées à partir de 85 articles de la littérature des sciences du sol, l'archéologie et la paléoclimatologie. Notre étude statistique est focalisée sur les sols minéraux. La répartition géographique des sites est donnée à la Figure 4:

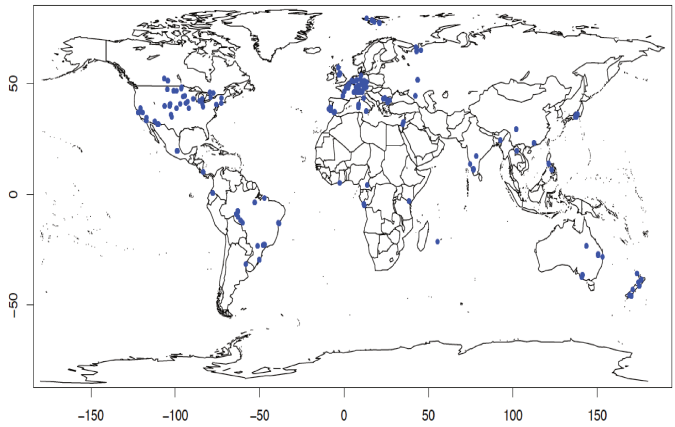


Figure 4: Localisation des sites étudiés (points bleus). La distribution spatiale des sites est plutôt hétérogène, avec un manque évident de données dans les zones extra-tropicales, en Afrique et en Russie.

Pour chaque profil, des informations géographiques (latitude, longitude et latitude), climatiques (température mensuelle, précipitations mensuelles, indice d'aridité, teneur en ^{14}C de l'atmosphère au cours de l'année de prélèvement, etc.) et environnementales (type d'écosystème et type de sol) ont été renseignées. Des mesures du radiocarbone et de la teneur en carbone ont été prises à différentes profondeurs.

Un nettoyage de la base de données à été effectué afin d'éliminer certains profils. Par exemple, on a éliminé les mesures réalisées sur les niveaux de litières et les profils dont les mesures ont été faites sur des molécules spécifiques ou des fractions granulométriques, densimétriques...non représentatives de la totalité de la matière organique du sol, etc.

Les facteurs climatiques et environnementaux potentiellement explicatifs de la dynamique, pris en compte dans cette méta-analyse statistique, ont été choisis en tenant compte des avis des experts et de façon à minimiser le problème de multicolinéarité. Cependant, on ne peut éviter qu'une dépendance naturelle existe entre les facteurs climatiques et environnementaux. Par exemple, au lieu de considérer les températures mensuelles, on a résumé cette information par la température moyenne annuelle et l'écart saisonnier de températures entre le mois le plus chaud et le mois le plus froid.

Les facteurs climatiques et environnementaux utilisés pour notre étude statistique sont résumés dans la Table 1:

Variables explicatives	Abréviation	Variables explicatives	Abréviation
Type d'ecosystème	Land	Type de sol	Soil
Précipitation annuelle moyenne	MAP	indice d'aridité	AI
Température annuelle moyenne	MAT	Décalage saisonnier de température	Dif_T
Précipitation mensuelle minimale	min_P	*$F^{14}C$ atmosphérique de l'année d'échantillonnage*	*$F^{14}Catm$*

Table 1: Facteurs climatiques et environnementaux explicatives potentiellement explicatifs de la dynamique du radiocarbone et de la teneur en carbone dans le sol. Le $F^{14}C$ atmosphérique de l'année d'échantillonage est mis en évidence en italique gras car il n'affecte que la dynamique du radiocarbone de sol.

Certains types de sols et d'écosystèmes sont regroupés afin de rendre le plan expérimental le moins déséquilibré possible. Un déséquilibre important de la base de données risque en effet de diminuer la précision de l'estimation.

Le chapitre de description de la base de données inclut aussi une discussion sur la répartition des profils du carbone des sols selon les différentes zones climatiques définies d'après la classification de Köppen. La base de données fournit une représentativité homogène du climat intermédiaire, tropical, chaud, tempéré et neigeux, en laissant de côté les climats extrêmes, aride et polaire (Figure 5).

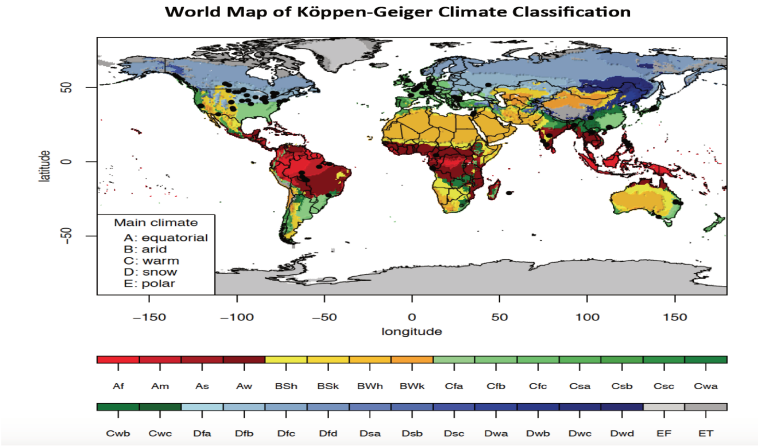


Figure 5: Superposition des sites sur la carte de classification climatique Köppen-Geiger. La "classification climatique de Köppen" est fondée sur l'analyse climatique de la période allant de 1951 à 2000. Elle divise les climats en cinq groupes climatiques principaux, chaque groupe étant divisé en fonction des précipitations saisonnières et des régimes de température. Les cinq groupes principaux sont A (équatorial), B (aride), C (tempéré chaud), D (neige) et E (polaire). Tous les climats, à l'exception de ceux du groupe E, sont ensuite déclinés en sous-groupés associés à des régimes saisonniers spécifiques de précipitations (représenté par la deuxième lettre). Par exemple, Af indique un climat de forêt tropicale humide. Le système assigne un sous-groupe de température pour tous les groupes, sauf ceux du groupe A, indiqué par une troisième lettre pour les climats B, C et D et une seconde lettre pour les climats E. Par exemple, Cfb indique un climat océanique avec des étés chauds comme indiqué par la dernière lettre b. Les climats sont classés en fonction de critères spécifiques propres à chaque type de climat.

La mauvaise répartition des années d'échantillonnage pour les sites avant la période de la bombe nucléaire (79%

des profils échantillonnés dans les années 1990) implique que la basse de données n'est pas représentative de la plage de variation du $F^{14}C$ atmosphérique surtout par rapport aux variations dues aux essais nucléaires atmosphériques.

Le nombre de mesures de radiocarbone et de la teneur en carbone varie entre 3 et 73 pour les profils échantillonnés. La majorité des profils de la base de données ont des enregistrements qui varient entre 4 et 10 observations. Ce nombre est satisfaisant pour avoir une idée sur la structure des profils du radiocarbone et de la teneur en carbone.

Une large variation naturelle est observée pour les teneurs en carbone en surface. Cette large variation est en accord avec la base de données mondiale sur le carbone du sol "SoilGrids".

Recherche stochastique de sélection de variables mixtes: application aux variables latentes du modèle hiérarchique de la dynamique du carbone des sols

Le deuxième chapitre du manuscrit est présenté sous forme d'un article publié le 13 septembre 2018 dans le Journal de la Société Française de Statistique (SFDS). Dans cet article, nous proposons une approche statistique bayésienne de sélection de variables pour mieux cerner la dynamique du carbone des sols en examinant la variation en profondeur du radiocarbone pour 159 profils sous différentes conditions de climat (température annuelle moyenne, précipitation annuelle moyenne, indice d'aridité, latitude, décalage saisonnier de température, $F^{14}C$ atmosphérique) et d'environnement (type de sol, type d'écosystème). Le modèle statistique utilisé dans cet article est inspiré du modèle statistique proposé par Mathieu et al. (2015).

Le modèle hiérarchique non linéaire à variance homogène d'occurrence locale des mesures (Figure 6), s'écrit de la façon hiérarchique suivante: pour un site $s \in [1:S]$, et pour une mesure $m_s \in [1:m_s]$, on modélise l'évolution du $F^{14}C$ du sol noté par $y(s,x)$ en fonction de la profondeur x par:

$$y(s,x) = \phi_1(s) + (\phi_2(s) - \phi_1(s)) \exp - \left(\frac{x}{\phi_3(s)} \right)^{\phi_4(s)} + \varepsilon(s,x) \quad \varepsilon(s,x) \sim N(0, \sigma^2)$$

- ϕ_1: *$F^{14}C$ en grande profondeur*.
- ϕ_2: *$F^{14}C$ en surface*.
- ϕ_3: *distance relative au point d'inflexion de la courbe*.
- ϕ_4: *décroissance plus ou moins forte*.

Les variables latentes ϕ_1, ϕ_2, ϕ_3 et $\phi_4(s)$ sont reliées linéairement aux variables potentiellement explicatives de la dynamique du radiocarbone: température annuelle moyenne, précipitation annuelle moyenne, $F^{14}C$ atmosphérique, indice d'aridité, décalage saisonnier de température, latitude, type de sol et type d'écosystème.

$$\begin{aligned} \phi_i &= X\beta_i + E_i \quad E_i \sim N(0, \sigma_i^2 I) \quad i = 1,2 \\ log(\phi_i) &= X\beta_i + E_i \quad E_i \sim N(0, \sigma_i^2 I) \quad i = 3,4 \end{aligned}$$

$\beta_i = (\beta_{i1}, \ldots, \beta_{iP})' \in \mathbb{R}^P$, où i = 1, 2, 3, 4, est le vecteur des effets de régression relative à la variable latente i, $E_i \in \mathbb{R}^P$ représente l'effet aléatoire désignant la variabilité inter-sites et $X \in M_{S,P}(\mathbb{R}))$ est la matrice de design construite en considérant un contraste traitement.

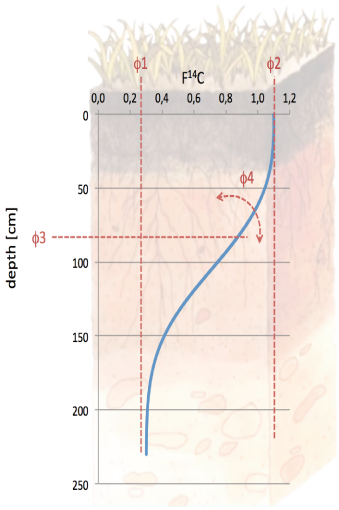


Figure 6: Profil statistique du $F^{14}C$ en fonction de la profondeur.

La recherche stochastique de sélection de variables (**SSVS**) est appliquée au niveau des variables latentes du modèle hiérarchique. 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. En fait, pour chaque variable explicative, un indicateur binaire I est associé permettant d'indiquer si cette variable est significative ($I = 1$) ou non ($I = 0$). Ainsi, une probabilité d'inclusion a posteriori est calculée à partir des itérations MCMC correspondant à I. Cette technique à été modifiée pour prendre en compte la sélection des variables explicatives qualitatives. Autrement dit, on affecte la même probabilité d'inclusion a priori pour toutes les modalités au sein de la même variable catégorielle. Le **SSVS** à été codé sous JAGS.

Nous discutons également de la performance pratique et des limitations de **SSVS** en présence de variables catégorielles et de la colinéarité entre certaines variables

En premier lieu, la performance du **SSVS** a été testée sur des données artificielles générées selon la structure du modèle statistique proposé pour les profils $F^{14}C$. Les résultats sur les données artificielles montrent que: 1- la présence de collinéarité augmente le taux de fausse détection du **SSVS** au niveau des variables latentes, 2- le **SSVS** peut ne pas détecter l'effet de certaines variables catégorielles significatives. Une analyse de sensibilité sur le choix de la probabilité d'inclusion a priori et sur la variance du prior Spike and slab sur les effets de régression a été également réalisée dans cet article.

Les meilleurs sous-modèles pour les variables latentes sont résumés dans la table 2:

Variables latentes	variables explicatives sélectionnées
$F^{14}C$ en profondeur (ϕ_1)	type d'écosystème, type de sol, température annuelle moyenne, décalage saisonnier de température
$F^{14}C$ en surface (ϕ_2)	type d'écosystème, $F^{14}C$ atmosphérique, température annuelle moyenne, décalage saisonnier de température, indice d'aridité
la profondeur d'incorporation $F^{14}C$ (ϕ_3)	type d'écosystème, type de sol, $F^{14}C$ atmosphérique température annuelle moyenne, précipitation annuelle moyenne, indice d'aridité, latitude, décalage saisonnier de température
Forme du profil $F^{14}C$ (ϕ_4)	type d'écosystème, latitude, $F^{14}C$ atmosphérique, température annuelle annuelle

Table 2: Les facteurs climatiques et environnementaux sélectionnés pour les variables latentes du modèle hiérarchique ϕ_1, ϕ_2, ϕ_3 et ϕ_4. Pour la troisième couche latente ϕ_3, toutes les variables explicatives ont été sélectionnées.

Les sous-modèles incluent toutes les variables explicatives détectées par le **SSVS** avec des probabilités

d'inclusion a posteriori au moins égales à 0.5 (Figure 7). De plus, pour être sûr que les variables catégorielles non détectées par le **SSVS** ne sont pas des fausses négatives, on a ajouté les variables catégorielles non détectées d'une manière successive afin de voir si une amélioration du critère DIC (Deviance Information Criterion) peut être établie.

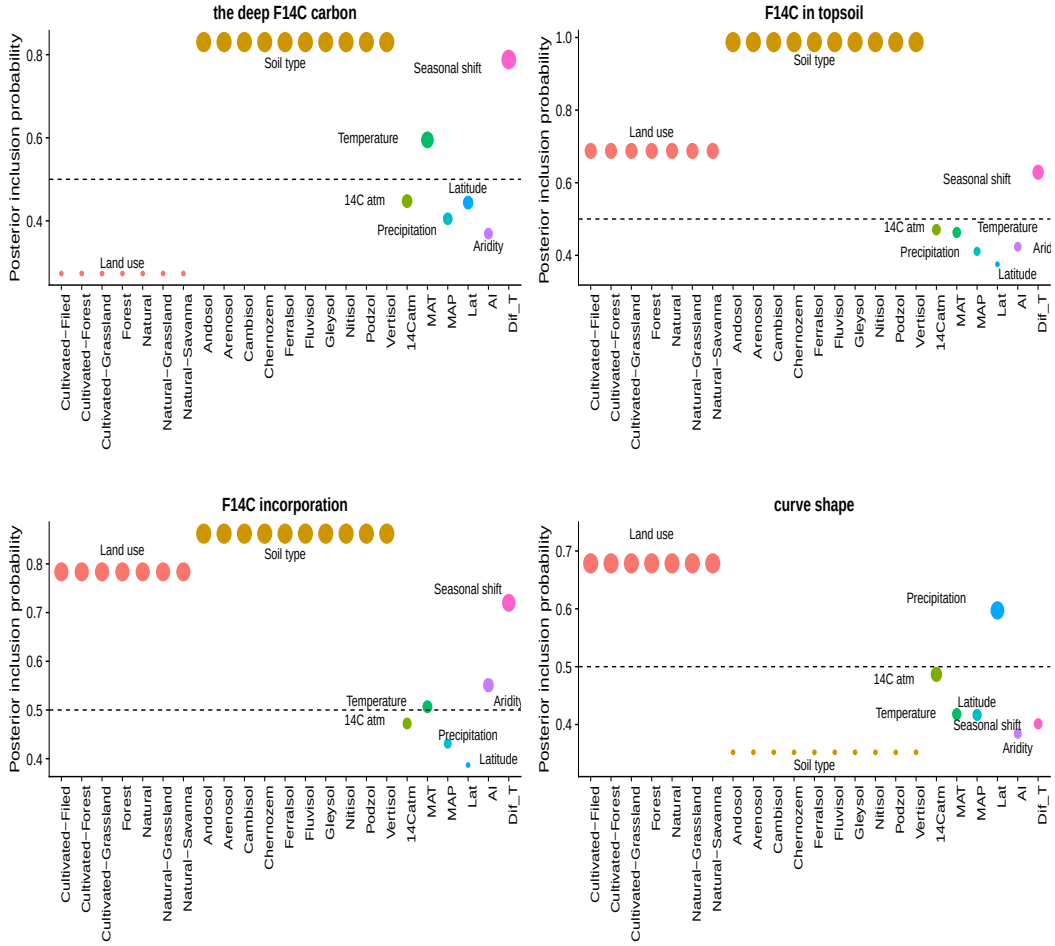


Figure 7: Probabilités d'inclusion a posteriori pour toutes les variables explicatives obtenues en appliquant la SSVS à aux profils $F^{14}C$ de la base de données. La taille des points dépend de l'importance de la probabilité d'inclusion a posteriori.

Exploration de trois méthodes bayésiennes de sélection de variables catégorielles et leurs codages sous JAGS

Le troisième chapitre de la thèse est présenté sous la forme d'un article, soumis le 15 août au journal Bayesian Analysis.

Le **SSVS** exploré dans le chapitre précédent peut conduire à faux négatifs pour certaines variables explicatives catégorielles. Afin de résoudre ce problème, un intérêt particulier a été porté à l'exploitation d'autres méthodes bayésiennes de sélection appropriées aux variables catégorielles.

Ce chapitre présente donc trois approches bayésiennes de sélection appropriées aux variables catégorielles. Parmi ces approches, le Bayesian Sparse Group Selection (**BSGS**) et le Bayesian Effect Fusion (**BEF**) permettent d'aller au-delà de la simple sélection des variables catégorielles. Le **BEF** peut être utilisé afin de fusionner les

modalités ayant le même effet au sein de chaque facteur et le **BSGS** nous permet d'établir un jugement probabiliste sur l'inclusion des effets des différentes modalités relatives au même groupe. Pour la derniére approche, le Bayesian Group Lasso with Spike and Slab (**BGL-SS**), l'estimateur *a posteriori* de la médiane présente une excellente performance, à la fois pour estimer et pour sélectionner les effets de régression. Notre objectif est d'appliquer ces techniques de sélection sur la couche latente du modèle non linéaire hiérarchique afin d'identifier les facteurs climatiques et environnementaux significatifs.

Plusieurs méthodes de sélection bayésienne ont été proposées dans la littérature afin de choisir le meilleur sous-modèle. On peut distinguer par exemple le Stochastic Search Variable Selection (SSVS) proposé par George and McCulloch (1993), la méthode de Kuo and Mallick (1998) ainsi que le Gibbs Variable Selection suggéré par Dellaportas and Ntzoufras (1997). Cependant, ces techniques capturent seulement les effets de régression relatifs à chaque variable continue et non pas les effets d'un regroupement des modalités associées aux variables catégorielles. Ici, on s'intéresse aux méthodes de sélection appropriées aux variables catégorielles qui exigent l'introduction de variables fictives (dummy variables) dans le modèle. Considérons d'abord le modèle d'analyse de la variance suivant:

$$Y = \mathbf{1}\mu + \sum_{g=1}^{G} X_g \beta_g + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \tag{1}$$

$(Y_1, \ldots, Y_n)$ est le vecteur réponse , G est le nombre de variables catégorielles et μ représente la constante associée au vecteur unitaire $\mathbf{1}$. Chaque variable catégorielle g renferme C_g modalités telles que $\sum_{l=1}^{G} C_l = p$. β_g est le vecteur des effets de taille $(C_g \times 1)$ relatif au facteur g. X_g est la matrice de design de taille $(n \times C_g)$ définie avec un contraste traitement et ε représente l'erreur.
Dans cet article, on présente et discute la performance de ces trois techniques de sélection appropriées aux variables catégorielles: Bayesian Sparse Group Selection proposé par Chen et al. (2016), Bayesian Group Lasso with Spike and Slab suggéré par Xu et al. (2015) et le Bayesian Effect Fusion using model-based clustering défini par Malsiner-Walli et al. (2017).

Spécification des priors:

- *Bayesian Group Lasso with Spike and Slab (**BGL-SS**)*:

$$\begin{aligned} \beta_g | \tau_g^2, \sigma^2 &\sim (1-\pi_g) N_{m_g}(0, \sigma^2 \tau_g^2 I_{m_g}) + \pi_g \delta_0(\beta_g) \\ \tau_g^2 &\sim G\left(\frac{m_g+1}{2}, \frac{\lambda^2}{2}\right) \\ \tau_g &\sim Ber(p_g) \\ \lambda &\sim G(a,b) \end{aligned} \tag{2}$$

Le **BGL-SS** est une technique qui permet d'estimer et de sélectionner les effets de régression simultanément. La formulation bayésienne du Lasso a été justifiée par Kyung (2010). Kyung et al. (2010) ont montré que le prior double exponentielle proposé par Tibshirani (1996) peut être écrit sous forme de la convolution d'une distribution gaussienne sur β_g avec un prior Gamma sur son paramètre d'échelle τ_g. Pour rendre le modèle plus *sparse*, Xu et al. (2015) ont considéré une loi de mélange entre une gaussienne et une masse de Dirac à 0 (pour avoir des effets qui valent exactement 0). Le résultat de la sélection est sensible au choix du paramètre de pénalité. Une petite valeur de λ tend souvent a préféré le modèle nul. Une valeur de 0.5 pour p_g $g = 1, \ldots, G$ est un choix raisonnable pour faire de la sélection. La règle de décision est basée sur l'estimation *a posteriori* de la médiane.

- *Bayesian Sparse Group Selection (**BSGS**)*:

$$\begin{aligned} \nu_g &\sim Ber(p_g) \\ \lambda_{lg}|\nu_g &\sim (1-\nu_g)\delta_0 + \nu_g Ber(p_{lg}) \\ \beta_{lg}|\lambda_{lg}\nu_g &\sim (1-\lambda_{lg}\nu_g)\delta_0 + \lambda_{lg}\nu_{lg}N(0,\tau_{lg}^2) \end{aligned} \tag{3}$$

  Le **BSGS** consiste à définir deux indicateurs binaires imbriqués ν_g (1: facteur g est sélectionné, 0: sinon) et λ_{lg} (1: la modalité l du facteur g est sélectionnée, 0: sinon). De plus, si le facteur g n'est pas sélectionné dans le modèle ($\nu_g = 0$), on affecte des indicateurs nuls pour toutes ces modalités. On peut poser $p_g = p_{lg} = 0.5$, ainsi, tous les sous-modèles sont équiprobables. δ_0 est une masse de Dirac en 0. Le choix de τ_{lg} a un effet important sur la sélection: par exemple une grande valeur de τ_{lg} pour $l = 1,\ldots,C_g$ diminue *a posteriori* la probabilité que le facteur g soit séléctionné. Cette technique nous permet d'avoir un jugement probabiliste non pas seulement sur l'inclusion des variables catégorielles dans le modèle mais aussi sur les effets des modalités au sein d'un même groupe.

- Bayesian Effet Fusion using model-based clustering (**BEF**):

$$\begin{aligned} \mathbb{P}(\beta_{gl}) &= \sum_{l=0}^{C_g} \nu_{cl} N(\beta_{gl}|\mu_l,\psi_g) \\ \nu_l &\sim Dir_{C_g+1}(e_0) \quad pour \quad l = 0,\ldots,C_g \\ \mu_0 &= 0 \\ \mu_l &\sim N(m_g, M_g) \quad pour \quad l = 1,\ldots,C_g \end{aligned} \tag{4}$$

  Cette approche est originale du fait qu'elle permet, non seulement de sélectionner les variables catégorielles significatives pour le modèle, mais aussi de fusionner les modalités au sein du même groupe ayant le même effet sur la variable réponse. Des détails supplémentaires sur le choix des hyperparamètres sont donnés dans l'article de Malsiner (2017). La règle de décision consiste à fusionner les modalités appartenant au même groupe de classification. Une variable catégorielle est éliminée du modèle si toutes ses modalités sont fusionnées avec la modalité de référence.

Les performances de sélection et l'analyse de sensibilité du réglage des hyperparamètres pour la spécification des priors ont été testées pour les trois approches de sélection dans une étude de simulation. Dans cet article, nous présentons également, en détail, la mise en œuvre des codes sous JAGS pour les trois méthodes de sélection bayésiennes.

Table 3: Comparaison de BGL-SS, BAGS et BEF, trois méthodes bayésiennes récentes appropriées aux variables catégorielles: quoi, comment et pour quoi faire?

		Méthodes bayéseinnes de sélection		
		BGL-SS	BSGS	BEF
pour quoi faire ?	Sélection des variables catégorielles	✓	✓	✓
	Sélection des modalités	✗	✓	✗
	Fusion des modalités	✗	✗	✓
Critère de sélection	probabilité d'inclusion a posteriori	✓	✓	✗
	médiane a posteriori des effets de régression	✓	✗	✓
	probabilité de fusion a posteriori	✗	✗	✓
Spécificité de la méthode	un nombre important de variables catégorielles et de modalités	✗	✗	✓
	prédicteurs avec un nombre important d'effets nuls	✗	✗	✓
	Rééquilibre du plan expérimental	✗	✗	✓
	Élimination totale de certaines modalités	✗	✓	✗
	Simplicité de la technique de sélection	✓	✗	✗
	Nécessite peu d'hyperparamètres à régler	✓	✗	✗
	jugement probabiliste sur la contribution des modalités	✗	✓	✗
	Non sensibilité à la probabilité d'inclusion a priori	✓	✗	✗

Applications des approches statistiques à la base de données

Le chapitre 4 de ma thèse est une application des techniques de sélection, détaillées dans les chapitres 2 et 3, sur les profils du radiocarbone et des teneurs en carbone de ma base de données.

Ce chapitre est composé de trois parties: la première partie compare toutes les méthodes bayésiennes de sélection introduites dans les chapitres précédents, la deuxième partie met en oeuvre l'interprétation physique des facteurs climatiques et environnementaux détectés comme significatifs, et finalement la troisième partie étudie le modèle statistique dans le cadre du changement global du climat et d'usage des sols.

En premier lieu, ce chapitre décrit les structures des modèles statistiques des profils du radiocarbone et des teneurs en carbone indépendamment. Une modification du modèle statistique des profils du $F^{14}C$ est mise en oeuvre par rapport au modèle statistique publié dans le journal SFDS. Par exemple, la latitude ne fait plus partie des variables explicatives et une loi normale tronquée est proposée pour modéliser la réponse $F^{14}C$.

Les meilleurs sous-modèles identifiés pour les profils de $F^{14}C$ et des teneurs en carbone

Les méthodes bayésiennes de sélection ont été comparées entre elles en se fondant sur des critères bayésiens de comparaison de modèles comme le *DIC*, l'erreur relative sur les sites de validation (cross-validation), etc. Les deux meilleurs sous-modèles obtenus après comparaison, sont résumés dans les tableaux suivants:

variables latentes	meilleur sous-ensemble de prédicteurs
$F^{14}C$ profond (ϕ_1)	type d'écosystème (Land) type de sol (Soil) température annuelle moyenne (MAT)
$F^{14}C$ en surface (ϕ_2)	type d'écosystème (Land) type de sol (Soil) précipitation annuelle moyenne (MAP) indice d'aridité (AI) décalage de température saisonnier (Dif_T)
Incorporation du $F^{14}C$ en profondeur (ϕ_3)	type d'écosystème (Land) type de sol (Soil) température annuelle moyenne (MAT) précipitation annuelle moyenne (MAP) indice d'aridité (AI) précipitation mensuelle minimale (min_P) décalage saisonnier de température (Dif_T)

Variables latentes	meilleur sous-ensemble de prédicteurs
Teneur en carbone profond (ω_1)	type d'écosystème (Land) type de sol (Soil)
Teneur en carbone en surface (ω_2)	type d'écosystème (Land) type de sol (Soil) précipitaion annuelle moyenne (MAP) température annuelle moyenne (MAT)
Incorporation du teneur en carbone (ω_3)	type d'écosystème (Land) type de sol (Sol) précipitation annuelle moyenne (MAP) décalage saisonnier de température (Dif_T)

Pour les profils $F^{14}C$, le meilleur sous-ensemble de facteurs climatiques et environnementaux est identifié par le Bayesian Group Lasso en se basant sur la probabilité d'inclusion *a posteriori*. En ce qui concerne les profils des teneurs en carbone, le meilleur sous-modèle est identifié à partir du Bayesian Effect Fusion, en se basant sur la probabilité de fusion *a posteriori* et la médiane des effets de régression.

Interprétations physiques des facteurs climatiques et environnementaux détectés comme significatifs

La deuxième partie de ce chapitre apporte une interprétation physique des facteurs climatiques et environnementaux détectés comme significatifs pour la dynamique du radiocarbone et des teneurs en carbone séparément.

Le radiocarbone atmosphérique, qui aurait dû ressortir parmi les variables explicatives du radiocarbone en surface selon les avis des experts, n'a pas été détecté comme significatif. Ce résultat peut être lié à la sur-représentation des profils échantillonnés en 1990 dans la base de données.

D'autre part, la température annuelle moyenne est détectée comme significative pour la teneur en carbone en surface, le $F^{14}C$ profond et l'incorporation du $F^{14}C$ en profondeur. Notre étude est en accord avec le résultat de Fang et al. (2005) qui montre que la matière organique non labile est plus sensible à la température que la matière organique labile.

Les précipitations annuelles moyennes influencent la signature en ^{14}C et la teneur en carbone en surface ainsi que l'incorporation de ces deux quantités en profondeur. Ces résultats peuvent être liés à la fois à la dilution des composantes des couches superficielles des sols par les composés organiques récemment introduits dans le sol (augmentation de la production primaire résultant de l'augmentation de la MAP) et au *priming effect* qui entraîne une perte des anciens composés organiques du sol.

D'autres interprétations physiques portent sur l'indice d'aridité et le décalage saisonnier de température .

Une surestimation des variables latentes correspondant à l'incorporation du radiocarbone et des teneurs en carbone en profondeur est identifiée. Cette mauvaise estimation, loin des valeurs qu'on peut avoir en réalité, peut être expliquée par la complexité du modèle et au lien non linéaire entre ces variables latentes et les réponses du radiocarbone et de la teneur en carbone.

Le résultat de la fusion des types de sol pour le radiocarbone en surface et pour celui en profondeur souligne que le profil de ^{14}C est davantage dominé par le climat/la végétation et la texture du sol pour les premiers centimètres du sol et par la teneur en argile pour les couches les plus profondes.

Prédictions du modèle statistique dans un contexte de changements globaux

L'avantage du modèle statistique est d'être utilisé pour prédire des profils en $F^{14}C$ et des teneurs en carbone dans des endroits où aucune donnée n'est disponible. Ici, en particulier on a essayé de prédire les profils du radiocarbone et de la teneur en carbone dans le cas de la conversion des forêt équatoriales en terres cultivées. Cette étude se base sur neuf profils localisés au Brésil. Une augmentation significative du radiocarbone en profondeur de 0.45 à 0.58, est observée pour les couches profondes. Autrement dit, le temps de résidence du carbone dans les couches profondes est plus long pour la forêt tropicale humide que pour les terres cultivées. Mais aucun changement n'est révélé pour la teneur en carbone en profondeur. Ces résultats sont conformes à ceux de Balesdent et al. (2018) qui montrent que l'utilisation des sols pour les cultures réduit l'incorporation de carbone dans la couche superficielle du sol, mais pas dans les couches plus profondes. Nos résultats ajoutent d'autres éléments à la discussion de Balesdent et al. (2018), allant au-delà du fait que la matière organique de nos sols actuels est l'héritage de sa gestion par plusieurs générations d'agriculteurs.

Par ailleurs, cette partie contient aussi une étude sur le reboisement des terres cultivées et prairies dans les

régions tempérés. Les résultats montrent que le reboisement des terres cultivées et des prairies tempérées entraîne une augmentation des stocks de carbone à court et à long terme.

Une étude aussi à été faite pour étudier l'augmentation de la température de 1, 1.5 et 2 °C sur la dynamique des profils du radiocarbone et celle de la teneur en carbone.

Conclusion et perspectives

Le dernier chapitre de ma thèse se divise en deux parties: la première résume brièvement les points principaux de la thèse partant du défi scientifique du départ et la mauvaise compréhension de la dynamique du carbone dans le sol jusqu'à l'avantage de la modélisation statistique.

La deuxième partie souligne les propositions et possibilités qui peuvent être examinées afin d'améliorer la structure du modèle statistique et d'optimiser les perspectives d'utilisation du modèle statistique.

Afin d'améliorer l'expression des incertitudes du modèle statistique, j'ai proposé un modèle de processus gaussien bivarié pour une modélisation conjointe du radiocarbone et de la teneur en carbone. En effet, notre modèle statistique ne prend pas en compte la dépendance entre les différentes mesures au sein d'un même site entre ces deux éléments. De plus, une visualisation de la variation du radiocarbone en fonction de la variation de la teneur en carbone met en évidence la présence d'une corrélation positive entre les deux réponses.

En ce qui concerne l'amélioration de la base de données, lors de l'analyse statistique, certains profils sortent nettement hors de la tendance générale. Ces horsains incluent notamment certains profils au Royaume-Uni sous un climat très humide. Ces profils ont probablement biaisé la sélection des variables et très probablement l'ampleur des diverses estimations. Une deuxième série d'évaluations peut être réalisée sans ces profils *aberrants* pour obtenir un modèle statistique qui mettrait mieux en lumière la tendance générale.

Un autre défi consistera à améliorer la base de données actuelle. Comme nous l'avions noté dans le chapitre 2, la base de données est loin d'être parfaite. Par exemple, nous n'avons pas le même nombre de profils par type de sol et d'écosystème. 37% des profils de la base de données sont des forêts, alors que seulement 8% sont définis comme des prairies cultivées. Même pour le type de sol, 9% des profils de la base de données sont définis en tant qu'Andosols et 7% en tant que Régosols / Arénosol / Leptosol. Cependant, seulement 1% de la superficie terrestre continentale de la planète est occupée par des Andosols et 22% par des Regosols / Arenosol / Leptosol. De plus, étant donné que le type de sol et l'écosystème sont associés, il pourrait également être intéressant de diviser Gleysol en deux catégories: les gleysols tropicaux et les gleysols boréaux. Certaines de leurs caractéristiques sont similaires car elles portent le même nom, mais d'autres, en particulier les interactions avec la végétation, sont différentes. L'augmentation du nombre de profils par catégorie de sol et d'écosystème (utilisation des sols + végétation) n'est toutefois pas le seul point important; il faut également tenir compte de plusieurs caractéristiques cruciales, telles que: l'occupation totale de la surface terrestre continentale par le sol et la couverture végétale, et l'association entre le sol, la couverture végétale et les conditions climatiques. De plus, la base de données actuelle ne contient pas de profils de classes climatiques arides et hyper-arides. C'est un manque crucial, en particulier pour l'utilisation du modèle en mode de prévision dans ces régions particulièrement vulnérables aux changements climatiques. Comme indiqué dans la description de la base de données, la répartition non homogène des années d'échantillonnage dans la base de données empêche une bonne représentation du profil $F^{14}C$, notamment pour la variable latente qui donne la profondeur d'incorporation.. Un autre point à améliorer dans la base de données est donc la distribution des profils par année d'échantillonnage. Dans la base de données actuelle, 53% des profils sont échantillonnés 1990 et 2000. En conséquence, le radiocarbone atmosphérique lié aux essais nucléaires au début des années 1960 n' a pas été

détecté comme significatif pour le $F^{14}C$ en surface ni pour l'incorporation du $F^{14}C$ en profondeur. Alors, pourquoi ne pas envisager l'analyse d'échantillons d'archives, comme ceux de Rothamsted? Cette collection d'échantillons a été créée par Lowes et Gilbert en 1843. Plusieurs milliers de sols recueillis dans les années 1920-1950 sont stockés dans le référentiel. Environ 1200 cultures et 200 échantillons de sol sont ajoutés chaque année aux archives.

L'extrapolation du modèle statistique bayésien développé pour la teneur en carbone des sols est utile pour obtenir une estimation globale (ou régionale) du stock de carbone des sols. Les modèles statistiques bayésiens pour la dynamique de la teneur en carbone et du radiocarbone nous permettent de prédire les profils du contenu en carbone et en carbone d'un nouveau site, en connaissant les informations climatiques et environnementales correspondantes. Comme aucune mesure n'est fournie pour ce site, les intervalles de crédibilité des paramètres inconnus du modèle seront plus larges que ceux observés pour les sites échantillonnés. En premier lieu, les profils prédits, lors du changement d'utilisation des sols ou des conditions climatiques, sont obtenus sans tenir compte des mesures observées (chapitre 5.5.2, section 5.5). Cela signifie que le site est considéré comme, un nouveau site véritable. Ainsi, de nouvelles variables latentes sont générées pour les modèles statistiques des profils $F^{14}C$ et de la teneur en carbone. Ces considérations suscitent la question suivante: comment prédire le profil $F^{14}C$ et de la teneur en carbone en fonction de l'évolution des conditions climatiques ou environnementales, en tenant compte des mesures déjà observées pour le site correspondant? Nous pouvons ajouter aux variables latentes actuelles et estimées, dans les conditions climatiques et environnementales actuelles des sites, le changement d'effet résultant du remplacement d'une forêt par une terre cultivée ou de l'augmentation de la température de 1 °C. En revanche, suivre cette proposition ne garantit pas la contrainte de positivité des variables latentes du modèle.

En second lieu, on peut se demander comment extrapoler le modèle statistique bayésien pour avoir un profil prédit de la teneur en carbone et du radiocarbone au niveau régional ou mondial. Avec une base de données plus complète et un dispositif expérimental plus équilibré, d'autres portes s'ouvrent. Il devient alors possible d'appliquer le modèle statistique conçu sur l'ensemble de la base de données à des modèles pour chaque type de sol et chaque zone climatique. Cela augmenterait la puissance de projection de l'étude. Cela permettrait de mieux déchiffrer l'impact du changement d'affectation des sols en fonction du type de sol et de mieux prédire l'impact du réchauffement climatique actuel selon les régions du monde. Se relier à un système d'information géographique (SIG) est également possible. On parle ici d'extrapolation à 3 dimensions: longitude, latitude et profondeur. La cartographie numérique des sols (DSM) utilisant des modèles spatiaux d'informations contextuelles de l'apprentissage profond (deep learning), est très populaire, et a déjà été utilisée pour générer des cartes (McBratney et al., 2003). En effet, il existe des méthodes d'apprentissage approfondi, telles que les réseaux de neurones à convolution, qui développent l'approche DSM classique en incluant des informations sur la proximité d'un site. Chaque site est caractérisé par des covariables climatiques et environnementales avec une matrice tridimensionnelle pour la largeur, la longueur en pixels d'une fenêtre centrée en un point (coordonnées du site) et en connaissant les covariables. L'apprentissage multitâche peut gérer la notion de profondeur en fournissant des prédictions, couche par couche. La possibilité d'extrapoler le modèle statistique serait très utile car l'optimisation de la conception de l'échantillonnage prend beaucoup de temps et est également coûteuse (acquisition de données et traitement des échantillons en laboratoire).

En outre, les approches de sélection bayésiennes peuvent aider à mieux comprendre les résultats du modèle mécaniste pour la dynamique du carbone des sols. Le coefficient de diffusion, qui traduit la bioturbation du sol, et le coefficient d'advection, lié à la diminution de la motilation, sont traités comme des constantes dans les modèles mécanistes développés pour la dynamique du carbone. Cependant, en réalité, ces coefficients ne sont pas constants et varient avec la profondeur. Un défi consisterait à transformer ces coefficients constants en fonctions, par exemple de type exponentielles décroissantes, de la profondeur. Les méthodes bayésienne de sélection explorées peuvent alors être utilisées pour définir les facteurs climatiques et environnementaux significatifs au sein de modèles statistico-mécanistes.

INTRODUCTION

Global warming is threatening the survival of human life and all life on earth. According to the latest assessment of the Intergovernmental Panel on Climate Change (IPCC) (Stocker et al., 2013), the average surface temperature of the planet has increased by nearly 0.9°C between 1901 and 2012. Global warming is mainly due to greenhouse gas emissions, in particular carbon dioxide, methane and nitrogen protoxide. The concentration of carbon dioxide has increased by 45% since the pre-industrial era (Harris, 2010) as a consequence of human activities that unbalance the global carbon cycle.

1.1- Soil plays a major role in the climate system

Soils are a major reservoir of carbon but its extent is not precisely assessed

The global carbon cycle has been extensively described and is a key factor in IPCC assessments. A comprehensive description is provided by Ciais et al. (2014) and is illustrated in Figure 1.1. In short, the ocean remains the largest carbon reservoir with 900 PgC for surface ocean [1] and 39,000 PgC for deep ocean, the atmosphere contains at present 828 PgC according to Prather et al. (2012) (about 590 PgC in pre-industrial times). The vegetation trap between 450 and 650 PgC (Prentice et al., 2001) and the soilcontains between 1500 and 2400 PgC in the form of organic matter (Batjes, 1996). As indicated by this large range used by the IPCC consortium, the extent of soil organic carbon is still poorly understood.

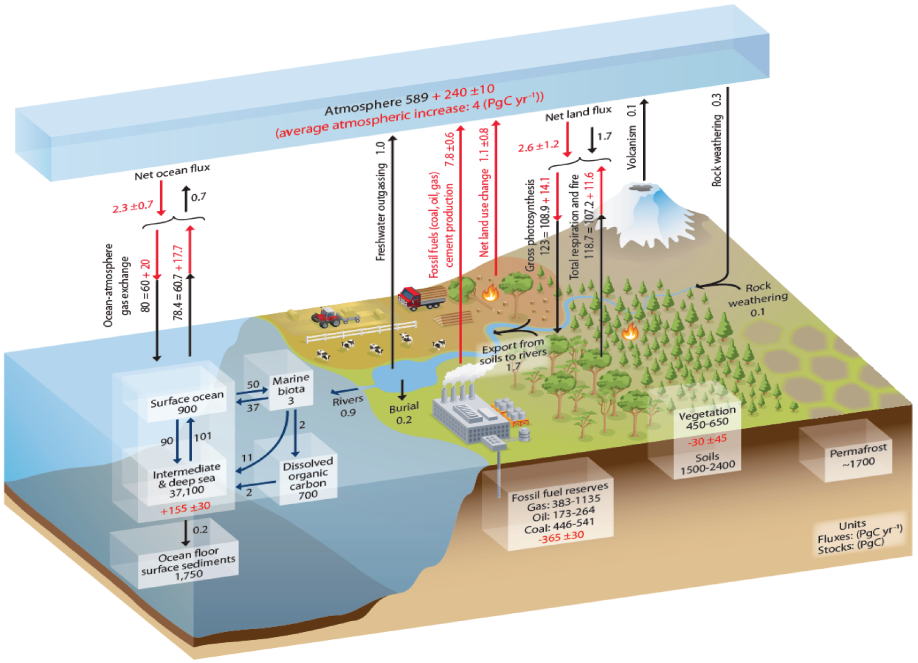


Figure 1.1: The global carbon cycle. Red arrows represent anthropogenic fluxes and black arrows represent natural fluxes in PgC yr^{-1} (Ciais et al., 2014).

Batjes (1996) underlined that the amount of deep soil carbon in particular is poorly known as few samples have been taken from the deeper layers.

Moreover, Batjes (1996) highlighted that the complexity of achieving a global estimation of the soil carbon pool stems from confounding the effects of climate, vegetation and land use on the one side and the lack of data for all soil types and climatic conditions on the other. In addition, a recent study by Tifafi et al. (2018) showed a great difference in regional and global soil carbon stock estimations based on a comparison between three global databases: the SoilGrids, the Harmonized World Soil Database (HWSD) and the Northern Circumpolar Soil Carbon

[1] A Petagram (Pg) is 10^9 tons, *i.e.* 10^{12} kg. It is also equivalent to 10^{15} grams.

Database (NCSCD) (Figure 1.2). The total soil carbon stock is estimated around 3400 PgC by SoilGrids, while it is about 2500 PgC according to the Harmonized World Soil Database.

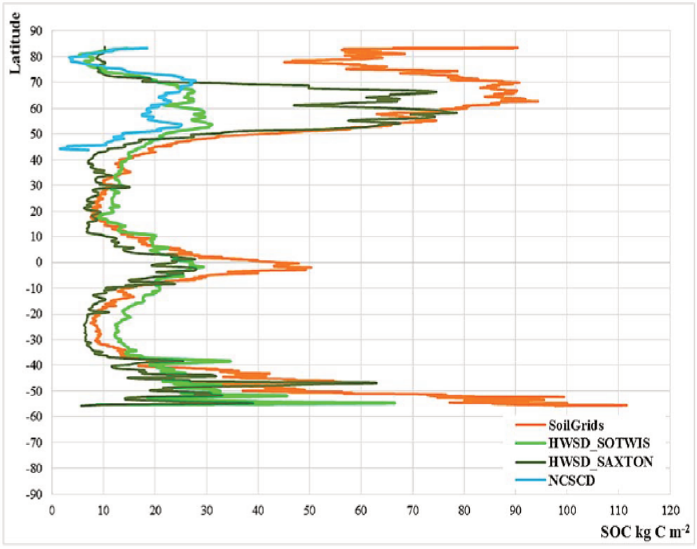


Figure 1.2: The total carbon stock (kg C m^{-2}) on the (0, 1 m) upper layer per latitude and for the different databases (Tifafi et al., 2018).

Whatever the precise figures, this reservoir is much larger than the one of the atmosphere and, with vegetation, is the only one that we can manage: increasing its capacity to stock more carbon would thus trap some of the fuel-derived CO_2 emitted.

Soils contribute the most to carbon exchanges with the atmosphere

The carbon cycle is a dynamic system with carbon fluxes from one reservoir to another. These fluxes are illustrated by arrows in Figure 1.1. Anthropogenic disturbance of the global cycle results from the increase in carbon fluxes to the atmosphere, mostly fossil fuels and cement production (emission of 7.8 $\pm$ 0.6 PgC yr^{-1}) and by land use change (emission of 1.1$\pm$0.8 PgC yr^{-1}). Vegetation traps approximately 123 PgC yr^{-1} of the atmospheric carbon via gross photosynthesis, of which 14.1 PgC yr^{-1} is from anthropogenic origin. Part of this carbon is introduced into the soil via plant roots and dead leaves while the other part is returned to the atmosphere via plant respiration : 107.2 PgC yr^{-1}. The net exchange between biosphere and atmosphere (60 PgC yr^{-1}, of which 49.4 PgC yr^{-1} for soil-atmosphere) is the largest exchange in the global carbon cycle.

Soil could help mitigate climate warming

As soil is a major reservoir of carbon and as net exchanges between soil and atmosphere are substantial, the IPCC (2013) highlighted the key role of soils as a part of climate change mitigation and adaptation.

The impact of land use changes was underlined in a study conducted by Deng et al. (2016). This meta analysis based on 103 recent publications for 160 sites in 29 different countries, showed that the soil carbon stock increases significantly after conversion from farmland to grassland (0.30 Mg $ha^{-1}yr^{-1}$) and forest to grassland (0.68 Mg $ha^{-1}yr^{-1}$), but declines significantly after conversion from grassland to farmland (0.89 Mg $ha^{-1}yr^{-1}$) and from forest to farmland (1.74 Mg $ha^{-1}yr^{-1}$).

The mitigation options are based on cropland management, grazing land management, and restoration of organic soils. These policies on agricultural practices and forest conservation, proposed to increase the soil carbon

uptake, have the advantage to better ensuring food security by preserving soil fertility. In that respect, a better understanding of the significance of the soil carbon pool was reached after the 2013 IPCC report as for the first time soils were considered as one of the resources for climate change mitigation. It is also worth mentioning the "Four per Thousand" initiative (https://www.4p1000.org) which aims at increasing the world soil carbon sequestration to a 40 cm depth at the rate of 0.4% per year in order to mitigate the global issues of climate change, food insecurity, and environmental pollution (Lal, 2016). To distinguish between sequestration and storage, it is commonly established that sequestration should be sustainable (at least 100 years, as recommended by the Kyoto protocol), whereas storage may be either short-term or long-term.

The large capacity of carbon exchanges with the atmosphere, the huge uncertainties about the response in soil carbon to global changes in climate and land use practices (positive or negative feedback) and lastly the fact that soil carbon is the only pool that humans can manage. All these factors show the crucial global interest of better understanding the fate of soil carbon.

1.2- The current representation of soil carbon dynamics in Land Surface Models is not entirely satisfactory

Current representations of mechanistic models for soil carbon dynamics

Several mechanistic models have been proposed to express the variation of soil carbon with depth. In these models, the representation of the physical processes at work is incomplete, however. Further research is therefore required to improve the simulation of soil carbon dynamics. The most widely used soil carbon models are included in the RothC Model (Coleman et al., 1997) which also incorporates a simple soil water model, and the CENTURY model system (Parton et al., 1987) which includes more complex models for plant growth, management operations, etc. The RothC model simulates profiles for the top 30cm of soil while the CENTURY model simulates profiles for the top 20cm (Falloon and Smith, 2010). Most mechanistic models are calibrated for the top centimeters but there is a need to include deep soil carbon into models, particularly since this stable carbon can be reintegrated into the global carbon cycle thanks to changes in climate and land use practices. In addition, a study done by Todd-Brown et al. (2013) shows that the majority of Earth System Models (ESMs) cannot reproduce grid-scale variation in soil carbon and may be missing key processes. Differences across soil carbon models included in ESMs are primarily due to differences in the estimation of Net Primary Product (NPP) and the parametrization of soil decomposition sub-models. The weakness and the limitation of soil carbon dynamics models comes also from the fact that these models are parametrized under specific management and climatic conditions. Furthermore, ESMs seldom consider depth carbon and even when they do so, discretization does not consider the changes in physical conditions and superposition results of box model layers. This overview of mechanistic models for soil carbon dynamics points out the importance of considering the total carbon of soils and of extending conceptualizations of processes to all scales of time and space.

Newly revealed processes and deep carbon are missing

In addition, there are large uncertainties about the processes that slow down mineralization and protect the organic matter in soil. Among these processes, one can distinguish: spatial inaccessibility to microorganisms and enzymes, hydrophobicity, encapsulation in organic macromolecules, litter resistance, organic matter-mineral interactions, etc. Until now, a major challenge has been to prioritize the role and impact of the stabilization process on soil carbon models (Paul, 2016). It will be a great challenge to express the new concepts of soil carbon stabilization/destabilization by differential equations in order to incorporate them into mechanistic modeling. Furthermore, the majority of soil carbon mechanistic models underestimate the amount of soil carbon since deep carbon is not considered in the

C-budgets Houghton (1995).

The incomplete view of the physical protection processes, the disregard of deep soil carbon layers and the parametrization of mechanistic models under specific management and climatic conditions highlight the need to better represent physically the simulation of soil carbon dynamics

Carbon isotopes provide clues to validate the representation of soil carbon dynamics

The best way to evaluate the performance of mechanistic soil carbon dynamics models is to compare them with data. This allows a direct comparison between field, lab, data and model outputs. In order to represent the organic matter, to specify it, to monitor processes and establish their kinetics, sample measurements of organic matter at several depths are needed. First, the amount of soil carbon can be defined by the carbon content produced by the lab analysis of measured samples. Secondly, there exist isotopic tracers methods using ^{13}C and ^{14}C to quantify the residence time of the natural organic matter in the soil for a few days to several thousand years.

The first tracing technique is based on monitoring the abundance of ^{13}C in the event of vegetation changes (from plant C3 to C4 or vice versa). Unfortunately, the data available for this technique are not sufficient for model evaluation and the technique requires a change in the type of photosynthesis performed by the vegetation.

The second technique, radiocarbon dating, is more powerful. The mean residence time of organic matter can be determined since the radiocarbon is characterized by its radioactive decay with a half-life of about 5730 years. In addition, the soil is considered as the memory of remarkable variations in radiocarbon activity, in particular the bomb peak due to atmospheric nuclear testing (Figure 1.3). The ^{14}C contents of the atmosphere doubled in 1962 compared to their natural level taken equal to the content of the atmosphere in the Northern hemisphere in 1950.

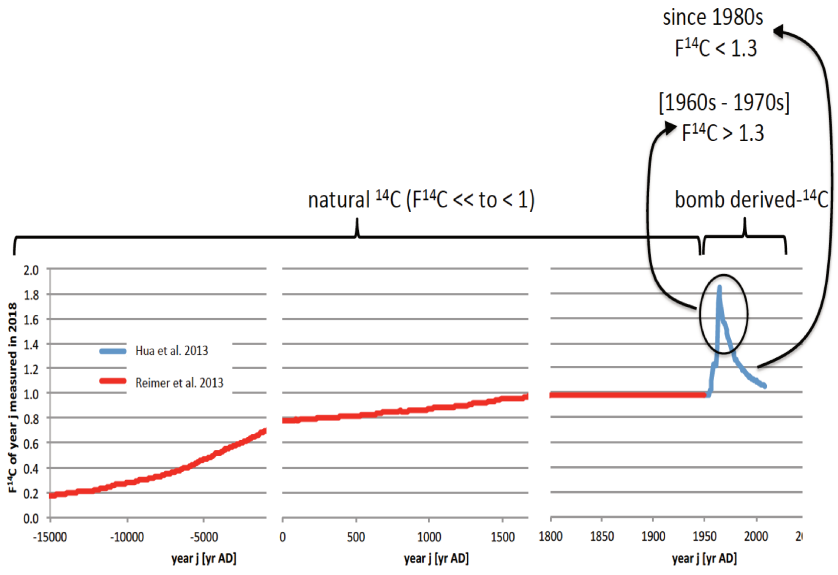


Figure 1.3: Amount of $F^{14}C$ of the year x measured in 2018, for any atmospheric material. The soil reproduces the variation in atmospheric radiocarbon as part of the radiocarbon reservoir but, by mixing organic material from several periods, the atmospheric signal is greatly smoothed in soil.

1.3- A statistical approach is an alternative to better represent uncertainties on soil carbon dynamics

In fact, integrating the new concepts of soil carbon protection into mechanistic models, considering what takes place in deep soil layers, and extending the conceptualization of processes to all scales of time and space, requires intensive development work. For that reason, a statistical approach based on observed empirical soil carbon data is useful in order to understand soil carbon dynamics, represent the different sources of uncertainties and provide answers about soil carbon behavior for the near future. The first attempt at the statistical modeling of soil carbon dynamics was made by Mathieu et al. (2015). Their proposed statistical approach was based on a world wide meta-analysis of 122 soil radiocarbon profiles collected from 87 articles in the soil science and archeology/paleoclimatology literature under different climatic and environmental conditions. The unknown model parameters were estimated relying on frequentist statistical inference. The study also sought to identify the effect of climatic and environmental factors on soil carbon dynamics. The analysis done by Mathieu et al. (2015) showed that the age of topsoil carbon was primarily affected by the climate and vegetation. In contrast, the results obtained on deep soil proved that the carbon content was impacted more by soil taxa than by climate. Moreover, they argued that the dependence on soil type points out the effect of other pedologic traits such as clay content and mineralogy. However, their interpretation was based on an expert analysis of the predictive results obtained without considering any statistical selection procedure to assess confidence about these judgements.

1.4- Contribution of my research work The crucial aim of my research is to improve the statistical model proposed by Mathieu et al. (2015) in order to better express the soil carbon dynamics, using Bayesian inference for estimation. This inference has the advantage of taking into account the uncertainties on the unknowns and made it possible to integrate into the statistical model the knowledge on soil carbon dynamics given by soil science experts (see Appendix 7.2 for further details). A subsequent goal was to put into practice a Bayesian selection approach in order to assign a probabilistic judgment and numerically quantify the respective contributions of climatic and environmental factors such as: land use, soil type, temperature, precipitation, aridity index, etc. on soil carbon dynamics. A particular concern is to predict the gain or loss of soil carbon by computing the carbon stock and residence time when changes in temperature or land use occur. Moreover, it will be useful to know which type of land use conversion can sequestrate more soil carbon and predict the soil carbon response if the temperature increases by 1.5°C or 2°C. Finally, we propose a Gaussian Bayesian model that considers jointly the soil carbon content and radiocarbon activity. This model takes into account, on the one hand, the correlation between soil carbon content and radiocarbon, and on the other, hand the correlation between depth measurements. This model is constructed in such a way as to provide information on soil carbon at deep layers. Our study will be useful to have an overview of the behavior of soil carbon dynamics in a context of global warming and will help make some decisions concerning land use practices.

The statistical modeling of soil carbon dynamics has several important advantages: a better representation of uncertainties on soil carbon dynamics, the presence of various tools that numerically quantify quantities of interest for soil scientists, and faster responses to the issues of today and the near future.

The manuscript is organized as follows: in the first chapter we detail the soil carbon database used for the study. We illustrate the heterogeneity of data sources, the available climatic and environmental information and the variety of sample sizes between different sites. The second chapter is an article published in the Société Française de Statistique SFdS journal in which we discuss the statistical model used on radiocarbon data. This chapter underlines the performance of Stochastic Search Variable Selection (SSVS) which is a Bayesian selection approach used as a first attempt to numerically quantify the climatic and environmental factors. The results obtained on artificial data

show that SSVS can misrepresent some significant categorical explanatory covariates such as soil type and land use. However, a cross validation test on real radiocarbon data, conducted on the statistical model including all climatic and environmental factors and the SSVS model, showed that the latter achieves a better prediction and adjustment level. Chapter 3 is motivated by the results obtained on the SSVS approach. It gives an overview of three recent Bayesian selection methods appropriate for categorical potential predictors: Bayesian Sparse Group Selection, Bayesian Group Lasso based on spike and slab priors and Bayesian effect fusion using model-based clustering. In this chapter, these three methods are applied on a simple regression model in order to better understand the functions and the characteristic features of each of the prior specifications. This chapter also includes a tutorial on these three Bayesian selection methods using Just Another Gibbs Sampler (JAGS) for Markov chain Monte Carlo (MCMC) simulations and the fourth chapter applies the Bayesian Selection methods to soil radiocarbon and soil carbon content dynamics. Chapter 4 intends to propose possible physical interpretation of the selected climatic and environmental factors that explain the shape of radiocarbon and carbon content profiles. These variables are the ones selected by the best subset by the Bayesian selection. It also provides a synthetic view of profiles shape under different climate. The predictive capacity of the model is also tested under two scenarios of land use change (deforestation in equatorial region, agricultural decline and reforestation in temperate region) and under 3 scenarios of global warming. The manuscript ends by a conclusion of my thesis work and with some propositions and perspectives in order to improve the soil carbon statistical modeling and the database.

CHAPTER 2

SOIL CARBON PROFILES DATABASE

2.1 Soil carbon profiles database

The description of the soil carbon database will mostly focus on what relates to radiocarbon data. Additional information, specific to carbon content will be provided at the end of the different sections if needed.

2.1.1 Original database

The database of 343 soil carbon profiles was collected from 85 articles in the soil science and archeology/paleoclimatology literature (Appendix 7.3). Permafrost and histosol were omitted from our survey, which focuses exclusively on mineral soil. Carbon in mineral forms is not considered, neither in this study nor in the database. So, in the rest of the document, "carbon" will be used for "organic carbon". The worldwide distribution of the 343 sites is illustrated in Figure 2.1:

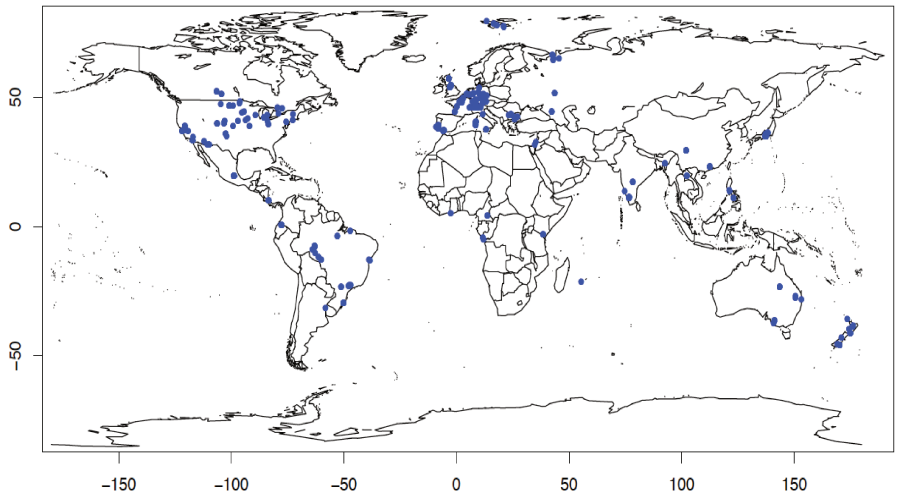


Figure 2.1: Geographical locations of studied sites are represented in blue dots. The spatial distribution of sites is very heterogeneous, with a lack of data in extra-tropical zones, Africa and Russia.

For each site, we collected the following information:

- Geographical coordinates: latitude and longitude (in decimal degrees), altitude (m). If not reported, the latitude, longitude, and altitude were determined from the site name with topocoding (http://www.topocoding.com).
- Climatic conditions: mean annual and monthly precipitation (mm), mean annual and monthly temperature (°C), aridity index. When not reported, the location coordinates were used to extract climatic values from the Climatic Research Unit (CRU) database (http://www.cru.uea.ac.uk/; New et al. (2002)). The mean annual aridity index, defined by the United Nations Environment Programme (UNEP) as the ratio of annual precipitation to annual potential evaporation, was obtained from the Food and Agriculture Organization of the United Nations (FAO) 10-minute mean climate grids for global land areas for the period 1950–2000 (Zomer et al., 2006) (Zomer et al., 2008). The index characterizes aridity for values ≤ 0.5, according to FAO.
- Soil type: as reported in the article, either according to the key FAO soil units http://www.fao.org/soils-portal/soil-survey/soil-classification/fao-legend/key-to-the-fao-soil-units/, the United States Department of Agriculture (USDA) https://www.nrcs.usda.gov or local classification. The descriptions were then assigned to large groups of soil types according to the World Reference Base (WRB) classification (Deckers and Nachtergaele, 1998).
- Land-use is reported as "natural" or "cultivated".

- Vegetation is heterogeneously reported: from vague information (*e.g.* "natural vegetation"), to very detailed information (including precision at the species level). In most cases, the main type of ecosystem (forest, field, grassland, savanna, desert) is also indicated.
- Soil levels: they are reported as upper and lower depths (cm) of the sampling slice.
- The soil fraction on which ^{14}C and organic carbon content were measured. It includes bulk, density fraction, particle size fraction, molecular fraction, even the specific molecule.
- Soil horizon (*i.e.* L for litter, O horizon, A, B, etc.) if available.
- Based on the soil horizon designation or designated as such in the article, paleosols are also specified.
- Radiocarbon activity provided for different sampling levels. Various units are used: yr BP, pMC, $F^{14}C$, $\Delta^{14}C$ (see Appendix 7.1). All values are reported as such and translated into $F^{14}C$.
- Soil organic carbon content provided for different sampling levels, as organic carbon concentration and/or as stock depending on what is available in the article. Soil organic carbon concentration is given as %wt or g/kg or derived unit. Soil organic carbon stock is expressed as kg/ha and derived unit.
- Bulk density is seldom available and if so, is given either for different depths or as the mean value whatever the depth.
- Other information such as clay content, granulometry, pH, soil texture are also reported when available (a few cases only).

2.1.2 Processed database

In order to focus on soil organic carbon in general and not on the specific aspect of dynamics, we removed the sites with the following features:

1. Soil levels corresponding to a "paleosol" (244 levels from 51 profiles) were removed since they no longer have any carbon exchange with the atmosphere.
2. Levels above the horizon O (soil litter) are not considered.
3. Some studies were carried out on specific molecules or granulometric fractions that are not representative of all soil organic matter. We only kept data obtained on a "bulk", "bulk after HCl" and a "bulk after concentrated HCl". These three supports provide a more correct overall picture of the total soil organic matter (152 profiles were removed).
4. Sites with overlapping layers were removed (this concerned two sites).
5. Thirty-four profiles with less than three observations were removed to ensure a good estimation and prediction for statistical inference for both soil radiocarbon and soil carbon content dynamics. In fact, four and three unknown parameters have to be estimated for the ^{14}C and organic carbon statistical models, respectively. Three observations is thus the minimum required number.
6. Profiles with an unknown soil type, land use or vegetation cover were removed: this concerned 8 profiles with unknown soil type and 32 profiles with unknown land use type.
7. Six profiles showing odd patterns of organic carbon distribution were removed from the carbon content dynamics modeling (but kept for ^{14}C modeling).

After data cleaning, only 131 profiles of soil radiocarbon and 125 profiles of soil carbon content remained for the statistical modeling.

Note on carbon content

To predict soil carbon profiles, we decided to work on soil carbon concentration (*e.g.* %wt) rather the soil carbon stock (*e.g.* g/m^2), since concentration is measured, while the soil carbon stock is calculated based on soil carbon concentration and on soil bulk density. However, bulk density is rarely or never provided in the articles used to build the database. Therefore, a pedotransfer function was used to predict the bulk density in the database, in order to complete the datasets. Alexander (1980) provided the most generic equation, where bulk density was derived from carbon concentration. But Alexander's equation is much too generic and does not account for soil type nor agro-pedo-climatic conditions, thus resulting in major uncertainties (Tifafi et al., 2018). Hence, although carbon stock is more relevant for agronomical and climatic purposes, in view of the fact that it would greatly increase the uncertainties compared to carbon concentration, it was decided to establish the statistical model on the carbon concentration profiles. A second step will be to extend to carbon stock, from the modeled profile.

2.1.3 Potential explanatory covariates affecting soil carbon dynamics

The behavior of soil carbon was investigated by modeling the dynamics of soil radiocarbon and of carbon content. Numerical (temperature, precipitation, etc.) and categorical (soil type, ecosystem type) predictors were considered for the meta-analysis. Explanatory covariates such as clay content, pH and granulometric information are not considered in this study since this information was seldom available.
The geographical information such as latitude, longitude and altitude are not considered in the statistical study since they do not impact the soil carbon dynamics. They are reflected in climatic parameters, such as temperature and to a lesser extent in "ecosystem".

2.1.3.1 Potential climatic numerical predictors

Climatic information is of prime importance to specify soil carbon dynamics. Statistically, taking all the monthly records of temperature and precipitation parameters (33 variables) into consideration would decrease the estimation and prediction performances of the linear model by increasing the variance of the estimated coefficients and making the model very sensitive to minor changes. In addition, it may enhance multicollinearity problems (Figure 2.2). For these reasons, in a first step, the number of predictors was reduced from 33 to 9. To select the potential climate predictors in this first step, we summarize information given by the monthly temperature and precipitation by considering:

- the extremes of temperature and precipitation regimes: minimum and maximum monthly precipitation (min_P and max_P, respectively), minimum monthly temperature (min_T and max_T, respectively),
- the mean annual temperature and precipitation (MAT and MAP, respectively),
- the seasonal shift between the warmest and coldest months (Dif_T),
- the seasonal shift between the wettest and the driest months (Dif_P),
- the aridity index (AI).

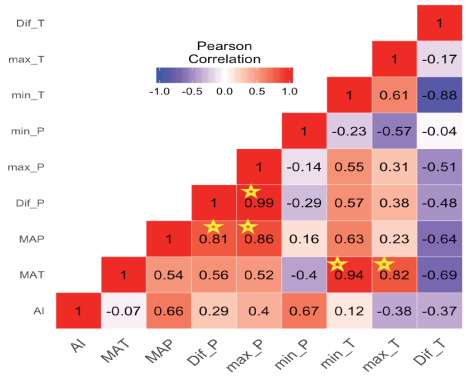


Figure 2.2: Correlation matrix corresponding to the original climatic numerical covariates for soil $F^{14}C$ profiles. The yellow stars highlight a strong collinearity between covariates. The more the correlation tends to -1 (dark blue) or +1 (dark red), the stronger the relationship between the two covariates. Both minimum and maximum temperatures are strongly correlated with the mean annual temperature, with Pearson Correlations (PC) of 0.94 and 0.82 respectively. The maximum precipitation and the seasonal shift precipitation are highly correlated with one another (PC = 0.99) and both are correlated with the mean annual precipitation (PC = 0.86 and 0.81 respectively).

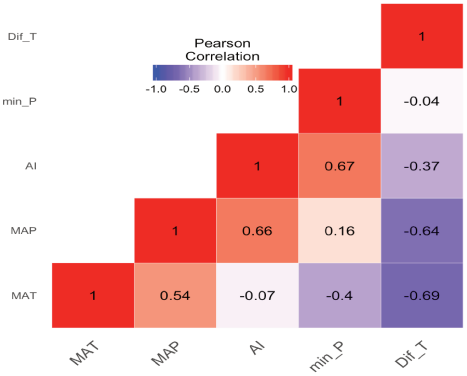


Figure 2.3: Correlation matrix after removing the strongly correlated covariates based on the $F^{14}C$ profiles.

To overcome the multicollinearity problem, the five climatic predictors chosen for the statistical modeling according to the Pearson correlation criterion are the following : MAT, MAP, min_P, AI and Dif_T (Figure 2.3).

2.1.3.2 Atmospheric ^{14}C concentration of the sampling year

Due to variation in the atmospheric ^{14}C content that derives from the bomb-peak (see Appendix 7.1), the representation of the soil radiocarbon dynamics is also affected by the atmospheric radiocarbon of the sampling year denoted by $F^{14}C$atm. This covariate is logically not used to express carbon content profiles.

2.1.3.3 Potential environmental categorical predictors : ecosystem and soil type

Soil type will be hereafter expressed according to its assignment to the large WRB group of soils. The remaining 131 profiles for radiocarbon statistical modeling were divided into 17 different soil types and 9 ecosystem types (vegetation + land use). To improve the balance of the experimental design and reduce the number of parameters, some soil types were grouped according to some shared physical properties, based on expert advice (see Table 2.1, Figure 2.4):

- Regosol-Arenosol, Arenosol and Leptosol were grouped together, as they share a lack of significant soil horizon.
- Kastozem and Phaeozem were pooled into Chernozem, as they are all humus-rich soils, at least for their surface layers.
- Fluvisol was merged with Cambisol, as they are often found in conjunction.
- Nitisol was merged with Ferralsol as they both contain a high amount of iron oxides that interact with organic compounds.
- Lastly, Plinthosol and Planosol were grouped with Gleysol as they are all susceptible to waterlogging and drought/frost stress.

WRB soil type group	profiles nb	merged WRB soil type group	short name	profiles nb
Arenosol Regosol-Arenosol Leptosol	3 1 3	"Areno-Regosol-like"	Areno-Regosol	7
Chernozem Kastanozem Phaeozem	16 1 2	"Chernozem-like"	Chernozem	19
Fluvisol Cambisol	1 15	"Cambisol-like"	Cambisol	16
Gleysol Planosol Plinthosol	7 1 1	"Gleysol-like"	Gleysol	9
Nitisol Ferralsol	4 14	"Ferralsol-like"	Ferralsol	18
Luvisol	27	Luvisol	Luvisol	27
Podzol	16	Podzol	Podzol	16
Vertisol	7	Vertisol	Vertisol	7
Andosol	12	Andosol	Andosol	12

Table 2.1: Merging of WRB soil type groups for soil radiocarbon profiles according to expert advice. For ease of reference, we will hereafter use the soil type group "short name" (*e.g.* Chernozem) to refer to the concatenation of the merged groups (*e.g.* Chernozem, Kastanozem, Phaeozem).

We grouped "land use" and "vegetation" into a single term, "ecosystem" that combines the two types of information. We identified 9 categories that we further merged into 6 groups (Table 2.2) : field, forest (forest, natural-forest), cultivated-forest, natural-grassland, cultivated- grassland and undefined natural (natural + natural-desert + natural-savanna) (see Table 2.2, Figure 2.4). The aggregation of ecosystem type was done in order to include the anthropogenic impact. In order to avoid categories with a small number of observations and to increase the prediction power of the statistical model, we created a group called "others" (Table 2.2).

ecosystem database	profiles nbr	short name for the merged ecosystem	profiles nbr
forest natural-forest	7 42	natural forest	49
cultivated-forest	10	cultivated forest	10
natural natural-savanna natural-desert	8 4 1	others	13
natural-grassland	33	natural-grassland	33
cultivated-grassland	8	cultivated-grassland	8
field	18	field	18

Table 2.2: Ecosystem aggregated types for soil radiocarbon profiles according to expert advice.

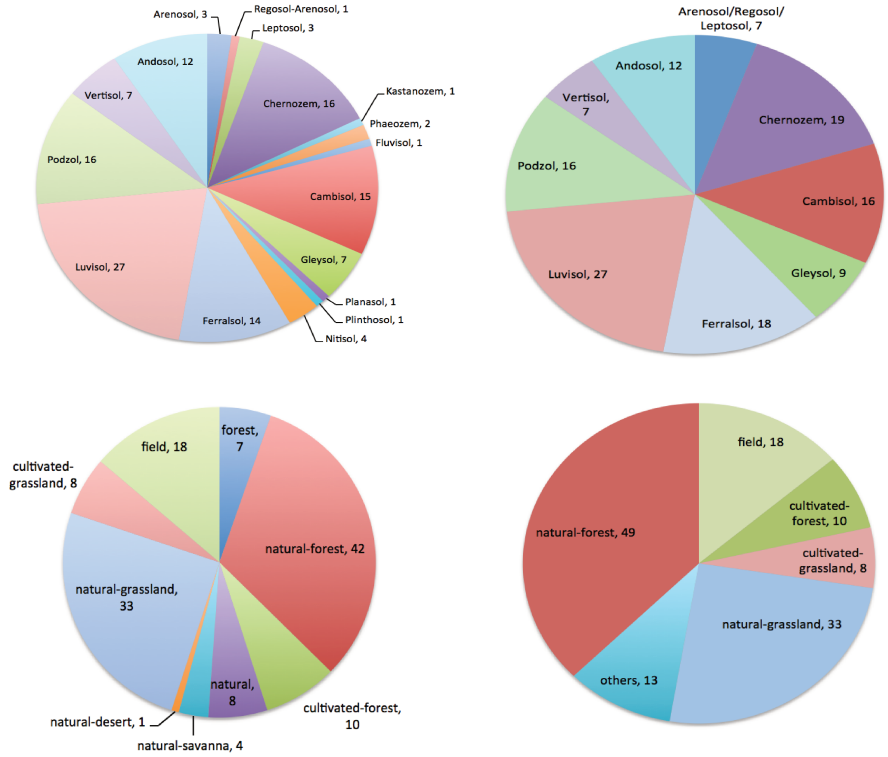


Figure 2.4: Graphical distribution of the 131 soil radiocarbon profiles before (left panels) and after (right panels) the merging of original ecosystem and soil type groups.

Note on carbon content

Note that the 6 profiles that differentiate between the 131 profiles for $F^{14}C$ modeling and the 125 profiles from organic carbon, are balanced as in Table 2.3:

nb of profiles	soil type
2	Chernozem
2	Gleysol
1	Ferralsol
1	Luvisol

nb of profiles	ecosystem type
4	natural-forest
1	others
1	field

Table 2.3: The categories of soil and ecosystem types for the six profiles removed for the soil carbon content statistical modeling.

Removing six profiles does not impact the experimental design shown in Figure 2.4. Four of the six removed profiles belong to the "natural-forest" ecosystem which includes the largest number of profiles. The same can be said for the soil type representativeness as we only removed one profile out of 18 Ferralsols and 27 Luvisols and 2 profiles out of 19 Chernozems. Only the removal of two profiles out of the nine Gleysols might affect the representativeness of the Gleysol category slightly.

2.1.4 Final selected climatic and environmental potential predictors for the soil carbon dynamics

The final climatic and environmental predictors used for the statistical meta-analysis are summarized in Table 2.4:

Potential covariates	Abbreviation	Potential covariates	Abbreviation
ecosystem type	Land	soil type	Soil
mean annual precipitation	MAP	aridity index	AI
mean annual temperature	MAT	seasonal temperature shift	Dif_T
minimum precipitation	min_P	*atmospheric $F^{14}C$ of the sampling year*	*$F^{14}Catm$*

Table 2.4: The potential explanatory climatic and environmental covariates that may impact soil radiocarbon carbon dynamics as well as soil carbon content dynamics. Atmospheric radiocarbon is highlighted in bold italics since it only impacts the radiocarbon soil dynamics.

2.1.5 Database evaluation

After the global presentation of the database, this section describes in greater detail how the design of the database can impact the performance of the statistical meta-analysis.

2.1.5.1 World climatic zones and distribution of the profiles

The representativeness of the database from the point of view of worldwide climatic diversity is illustrated by the superimposition of the profile locations on the Köppen-Geiger climate map (Kottek et al., 2006) (Figure 2.5) and by the cumulative charts of the number of profiles according to climatic parameters (Figure 2.6).

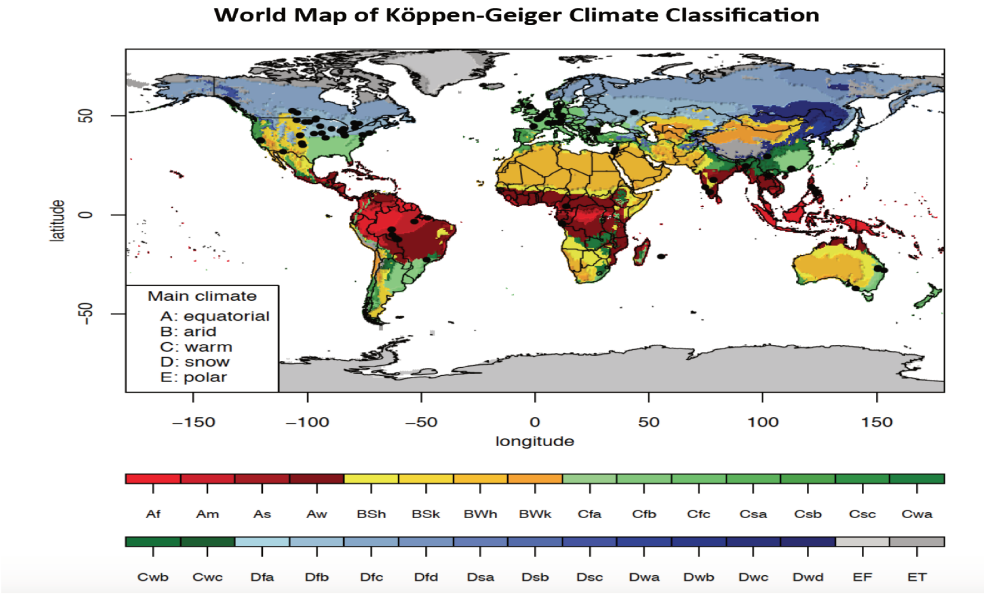


Figure 2.5: Superimposition of site locations on the Köppen-Geiger climate classification. The "Köppen climate classification" is based on the 50-year period 1951-2000. It divides climates into five main climate groups, with each group being divided based on seasonal precipitation and temperature patterns. The five main groups are A (equatorial), B (arid), C (warm temperate), D (snow), and E (polar). All climates except for those in the E group are assigned to a seasonal precipitation subgroup (represented by a second letter). For example, Af indicates a tropical rainforest climate. The system assigns a temperature subgroup for all groups other than those in the A group, indicated by a third letter for climates in B, C, and D, and a second letter for climates in E. For example, Cfb indicates an oceanic climate with warm summers as indicated by the ending b. Climates are classified based on specific criteria unique to each climate type.

Köppen published his first climate classification in 1884 (Köppen, 1884) and later improved it (*e.g.* Koppen (1936); Geiger (1954)) to achieve the **Köppen-Geiger climate classification**. It has since been regularly updated. Kottek provided the latest version in 2006 (Kottek et al., 2006) for the second part of the 20^{th} century. The classification is based on the division of earth climates into 5 main climate groups: equatorial (Figure 2.5: acronym beginning with A, red-derived zones), arid (B, yellowish), warm temperate (C, green), snow (D, blue), polar (E, gray). The main groups are then subdivided into subgroups (29 in total). The definition of climatic groups is provided in Kottek et al. (2006), table 1 and the first subdivision is reproduced as follows (Table 2.5):

Type	Description	Criterion	profiles nb
A	**Equatorial climates**	**$T_{min} \geq 18°C$**	**20**
Af	Equatorial rainforest, fully humid	$P_{min} \geq 60mm$	0
Am	Equatorial monsoon	$MAP \geq 25 (100 - Pmin)$	10
As	Equatorial savannah with dry summer	$P_{min} \leq 60mm$ in summer	4
Aw	Equatorial savannah with dry winter	$P_{min} \leq 60mm$ in winter	6
B	**Arid climates**	**$Pann \leq 10\ P_{th}$**	**1**
BS	Steppe climate	$P_{ann} \geq 5\ P_{th}$	1
BW	Desert climate	$P_{ann} \leq 5\ P_{th}$	0
C	**Warm temperate climates**	**$-3°C \leq Tmin \leq +18°C$**	**70**
Cs	Warm temperate climate with dry summer	$P_{Smin} < P_{Wmin}$ $P_{Wmax} \geq 3\ P_{Smin}$ $P_{Smin} \leq 40$ mm	5
Cw	Warm temperate climate with dry winter	$P_{Smin} \geq P_{Wmin}$ $P_{Smax} \geq 10\ P_{Wmin}$	18
Cf	Warm temperate climate, fully humid	neither Cs nor Cw	47
D	**Snow climates**	**$T_{min} \leq -3°C$**	**38**
Ds	Snow climate with dry summer	$P_{Smin} \leq P_{Wmin}$ $P_{Wmax} \geq 3\ P_{Smin}$ $P_{Smin} \leq 40$ mm	15
Df	Snow climate, fully humid	neither Ds or Dw	19
E	**Polar climates**	**$T_{max} \leq 10°C$**	**2**
ET	Tundra climate	$0°C \leq T_{max} \leq 10°C$	2
EF	Frost climate	$T_{max} \leq 0°C$	0

Table 2.5: Description of the Köppen-Geiger classification (1st and 2nd letter description only) and number of radiocarbon profiles selected from the database that correspond to the different subgroups (last column). P_{min} (P_{max}) and T_{min} (T_{max}) are for the minimum (maximum) monthly precipitation and temperature, P_{ann} is for the MAP, S and W subscripts are for summer and winter respectively. $P_{th} = 2 * MAT + a$, with $a = 0$ if at least $2/3$ of MAP occurs in winter, $a = 28$ if at least $2/3$ of MAP occurs in summer and $a = 14$ otherwise. The calculation key implies that the polar climates (E) have to be determined first, followed by the arid climates (B) and subsequent differentiations into the equatorial climates (A) and the warm temperate and snow climates (C) and (D), respectively.

Examination of the database shows that 20 of the selected profiles belong to "equatorial climates", 1 to "arid climates", 70 to "warm temperate climates", 38 to "snow climates" and 2 to "polar climates (Table 2.5). At the first order, this **results in a homogeneous representativeness of intermediate climate types, *i.e.* tropical, warm temperate and snow climates**, leaving out extreme climates. "Arid climate" is represented by only one profile from the Sonora Desert, AZ, USA) and "polar climate" by two Italian mountain profiles. **Warm temperate climates are overrepresented**, and this tendency is even stronger when compared with the land surface ratio they occupy (Figure 2.5). This is due to the fact that most agronomical studies have traditionally been performed in temperate regions, while investigations in other regions are a recent phenomenon. **At the second order, however, not all sub climates are present in the database**. Some sub-climates are overrepresented. So, whereas equatorial climates are well balanced between "monsoon" and "savannah climates", "rainforest climate" is not represented. The high weight of "fully humid warm temperate climates" is in line with the respective weight of the "dry season" and "fully humid" within this type of climate. The imbalance is rather between Cs and Cw where one would have expected an

equivalent weight, whereas in fact Cw are three to four times more abundant than Cs in the database. The same applies for snow climates, with an overrepresentation of the "snow climate with dry summer".

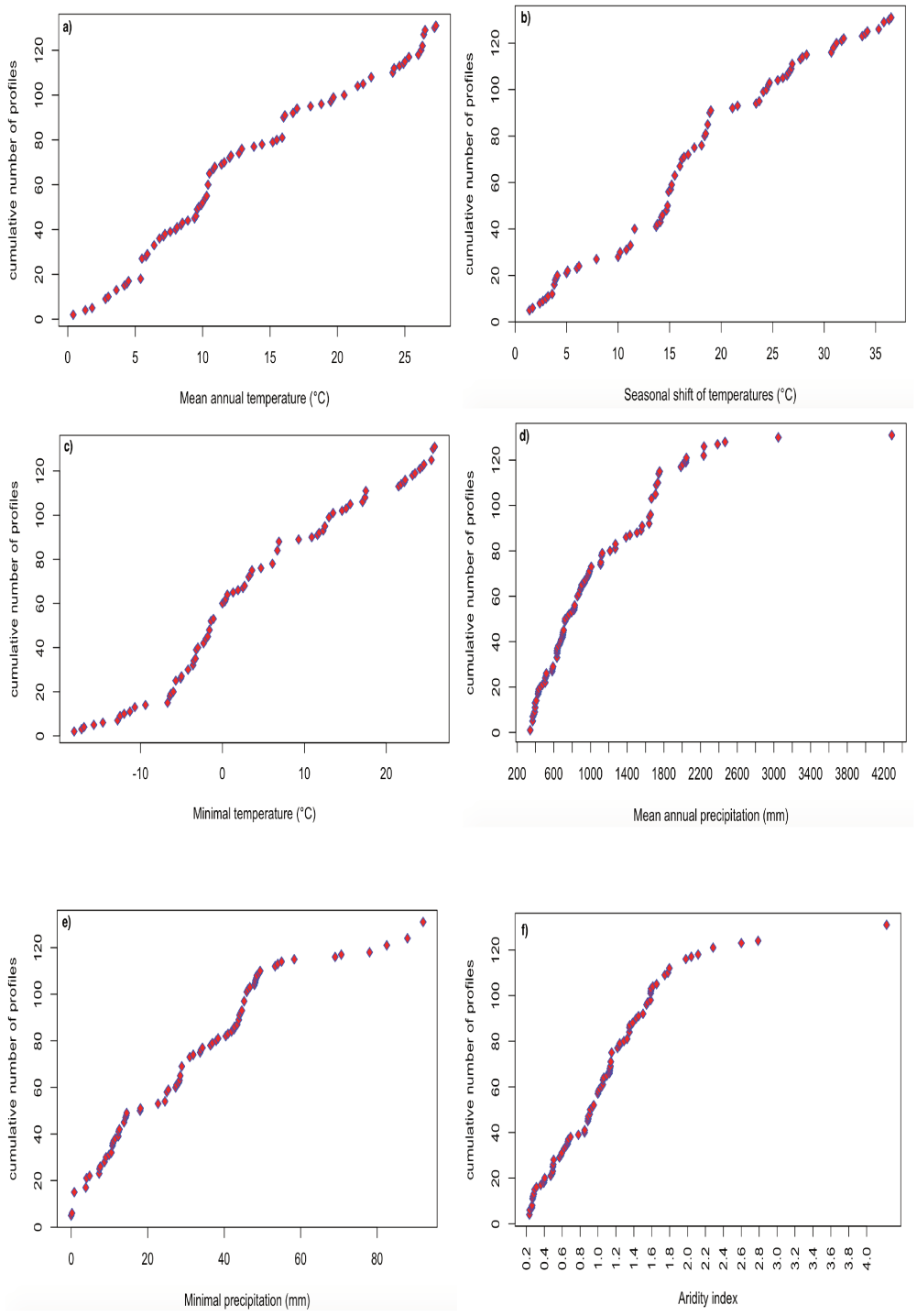


Figure 2.6: The cumulative number of $F^{14}C$ profiles for the following six numerical climatic covariates: mean annual temperature (panel a), seasonal shift of temperature (panel b), minimal temperature (panel c), mean annual precipitation (panel d), minimal precipitation (panel e) and aridity index (panel f).

Analysis of the annual climatic parameters (mean annual temperature, mean annual precipitation and aridity index) is shown in Figure 2.6. The database provides a relatively **homogeneous representation of both mean annual temperatures within the [1;28°C] range and mean annual precipitation within the [100;1800mm] range**, including also a dozen profiles from 2000 to 4000mm a year. The aridity index shows a similar representa-

tiveness with most of the profiles regularly distributed in the [0.15;2] range, *i.e.* from very arid to humid zones, and including 7 profiles above 2.

Temperature seasonal shifts vary from a few °C to more than 35°C (Figure 2.6, panel b), covering the wide range of climates from tropical to continental. The profile distribution within this range is less continuous than for annual parameters, and rather patchy. Nevertheless, **the [2; 38°C] range of seasonal shift is homogeneously swept.**

Briefly, our database will allow the statistical model to predict soil profiles under equatorial, warm temperate and snow climates but precludes using the model in extreme conditions such as arid climates, polar climates and very wet regions.

2.1.5.2 Soil type diversity and distribution of profiles

The database offers a good representativeness of the land coverage diversity. As shown in table 6, the 17 soil types from the database represent about 71% of the land coverage. However, the distribution between the 17 categories (or between the 9 merged groups of soil types) is unbalanced. So with 12 profiles Andosol, which only represents 1% of total land area on Earth will be much better constrained than areno-regosol represented by only 7 profiles but covering 22% of land area. The difference in representativeness is not that large for the other soil types.

WRB soil type	% of the total land area	merged WRB group of soil type	% of the total land area	nb of $F^{14}C$ profiles
Arenosol Regosol-Arenosol Leptosol	7 2 13	Areno-Regosol	22	7
Fluvisol Cambisol	2.8 12	Cambisol	14.8	16
Nitisol Ferralsol	1.6 6	Ferralsol	7.6	18
Gleysol Planosol Plinthosol	5.7 1.0 0.5	Gleysol	7.2	9
Chernozem Kastanozem Phaeozem	1.8 3.7 1.5	Chernozem	7	19
Luvisol	5	Luvisol	5	27
Podzol Vertisol Andosol	4 2.7(*)-2($^+$) 1	Podzol Vertisol Andosol	4 2.7(*)-2($^+$) 1	16 7 12

Table 2.6: Percentage of the total continental land area on Earth by soil types (first two columns) and by merged groups of soil type (columns 3 and 4). The last column gives the number of profiles by merged group. Note for Vertisol: the total land area differs according to the classification, values according to the FAO [*] and USDA [$^+$] classifications are provided Source : https ://www.britannica.com

2.1.5.3 Ecosystem diversity and distribution of profiles

The database offers a correct representativeness of ecosystems (Table 2.2). Each of the 6 selected categories is represented by 8 to 49 profiles, *i.e.* by 6 to 37% of the available profiles. However, it remains unbalanced with a high weight (45% of the selected database) of forests, mostly natural forests, whereas grasslands account for 31% leaving the remaining 24% for both fields and other types of ecosystems.

2.1.5.4 Sampling year and distribution of profiles

The distribution of sampling years over the last 60 years is poor in the database, with most of the profiles of the database sampled in the 1990s (79% - Figure 2.7). This means that the evolution of the bomb peak record by soil over the years is badly represented. An evolution of the $F^{14}C$ profile is expected according to the sampling year. Incorporation of the bomb peak, which can be more or less fast, is expected to be recorded by the upper layers of soil. The incorporation rate and the incorporation depth that should differ from one profile to another might not be captured in our dataset.

Because of the overrepresentation of the 1990s, we might not be able to perfectly mimic ^{14}C profiles. Profiles from the early 1960s with the highest expected $F^{14}C$ for the upper levels are scarce, as are profiles prior to the bomb peak. Atmospheric $F^{14}C$, which refers to the sampling year, will likely not consider the real variation of atmospheric radiocarbon with time. Nevertheless, as $F^{14}C$ is only a way to represent the carbon dynamics but does not impact it, this database unbalanced will not impair our definition of the impact factors nor the evaluation of their impact magnitude.

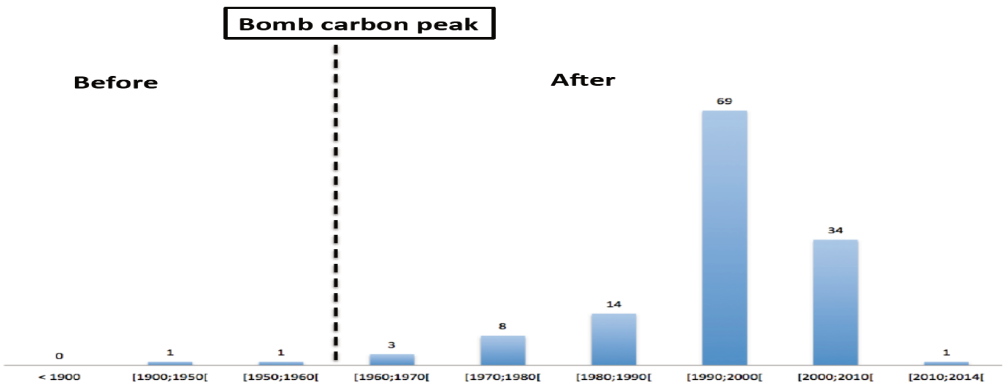


Figure 2.7: Distribution of profiles according to the sampling year (grouped by decade). The majority of the $F^{14}C$ profiles were sampled between 1990 and 2000. Only two profiles were collected before the "bomb peak".

In contrast, the misrepresentation of the sampling years will not affect the soil carbon content statistical modeling.

2.1.5.5 Variation in the number of observations per profile

Each of the $F^{14}C$ and carbon content measurements corresponds to a specific level characterized by an upper and a lower level depth.

The number of observations per profile is sufficient to assess the profile shape

The number and the distribution of observations within the profile are very heterogeneous (Figure 2.8). Their number varies from 3 to 73 observations. For example, an Indian Luvisol profile under a cultivated field sampled by Becker-Heidmann and Scharpenseel (1989) at a constant 2cm sampling step to 158 cm depth (Figure 2.9, top panel b), yielded 73 observations whereas 31 profiles have only 3 sampled layers of different thickness. The majority of the soil carbon profiles correspond to a number of measurements between 4 and 10 observations. Four observations are sufficient to constrain the $F^{14}C$ profile shape. The distribution of the number of observations between the profiles remains similar if we look at the level of the type of soil or type of ecosystem (Tables 2.7 and 2.8 respectively). All the categories mainly contain profiles with 4 to 10 observations.

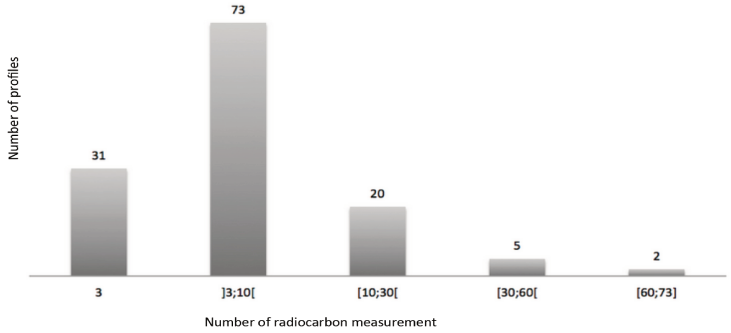


Figure 2.8: Distribution of $F^{14}C$ profiles according to the number of measurements. Most sites have between 3 and 10 measurements.

Soil types	3	[4;10[	[10;30[	[30;60[	[60;73[
Andosol	3	5	1	2	1
Arenosol/regosol	1	3	3	0	0
Cambisol	8	7	1	0	0
Luvisol	7	17	2	1	0
Ferralsol	3	9	6	0	0
Chernozem	4	11	4	0	0
Podzol	4	9	1	1	1
Vertisol	1	5	1	0	0
Gleysol	0	7	1	1	0

Table 2.7: Number of $F^{14}C$ profiles according to the number of measurements for each aggregated soil type.

ecosystem type	3	]4;10[	[10;30[	[30;60[	[60;73[
cultivated-field	10	8	0	0	0
cultivated-forest	5	4	1	0	0
cultivated-grassland	0	6	1	1	0
natural-forest	8	29	8	3	1
natural-grassland	7	20	6	0	0
others	1	6	4	1	1

Table 2.8: Number of $F^{14}C$ profiles according to the number of measurements for each aggregated ecosystem.

4 to 10 observations avoids giving too much weight to particular points without losing the specific structure of the profiles

The sampling can be either continuous, *i.e.* all the profile is sampled through several successive layers (Figure 2.9 and c) or discontinuous, *i.e.* some slices here and there in the profile (Figure 2.9). The sampling step can differ for $F^{14}C$ and for carbon content. For example, a 2cm step for $F^{14}C$ (Figure 2.9 top panels) and a 40cm step for carbon content (Figure 2.9, bottom panels). In such a case, because the database was primarily designed to gather $F^{14}C$ data, the levels are those defined by the $F^{14}C$ data distribution. So, the measured carbon content value for the large carbon content layer is repeated for all the small $F^{14}C$ levels included in it. Otherwise, when the $F^{14}C$ layer is larger than the one for the carbon content, a mean value of carbon content is reported as carbon content in the $F^{14}C$ layer. **This choice, dictated by the structure of the database, in the case of different sampling steps between $F^{14}C$ and carbon content profiles, may bias the original signal.** Fortunately, this only concerns 9 profiles from the database.

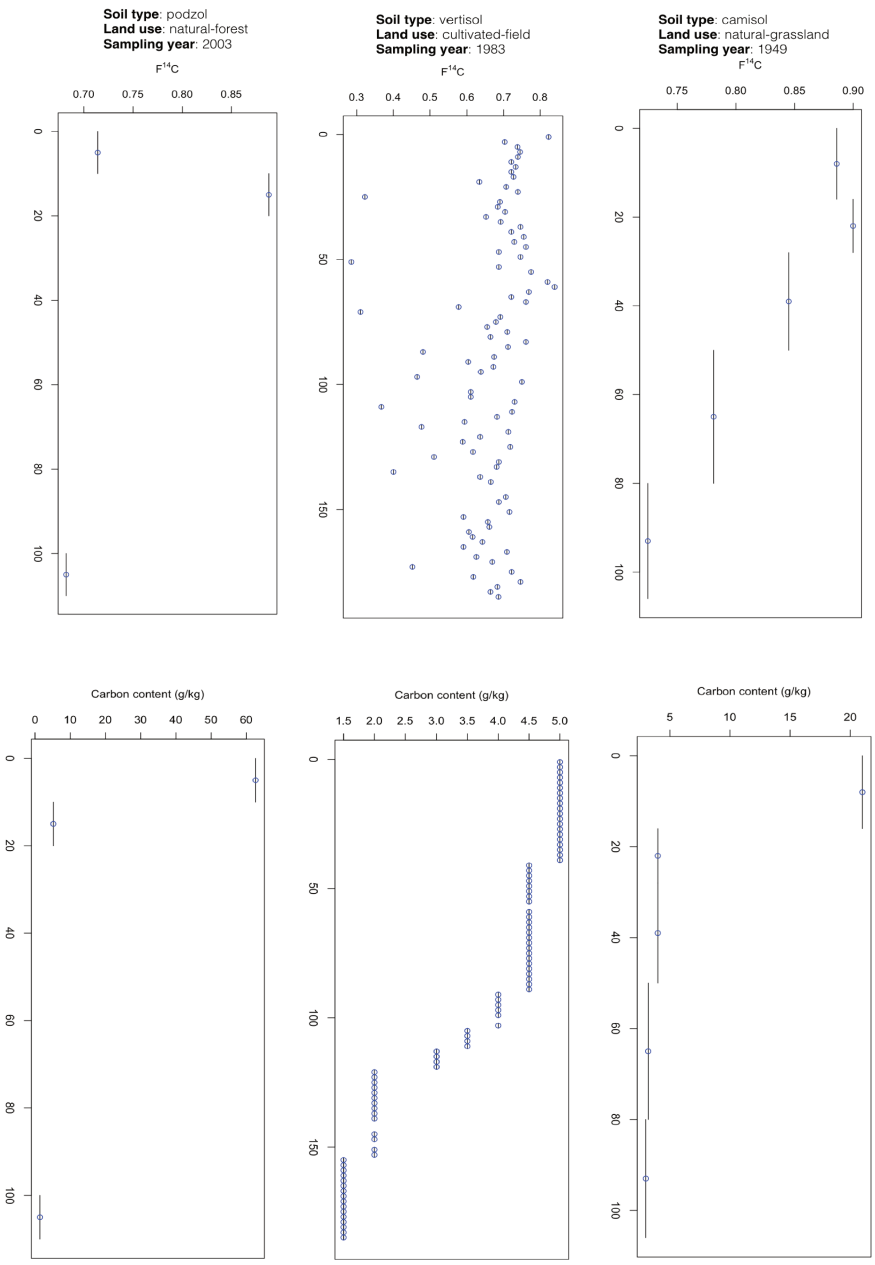


Figure 2.9: The $F^{14}C$ (upper panels) and carbon content (lower panels) profiles for three profiles from the database: discontinuously sampled (panels a), continuously sampled (panels b and c) with a different (panels b) and the same (panels c) sampling step for both $F^{14}C$ and carbon content (panels c). The black lines refer to the soil layer from which the radiocarbon and carbon content were sampled. The blue points indicate the mean level corresponding to the sampled soil layers.

2.1.5.6 Large variation of the topsoil carbon content

The large variability of carbon content in the database is in agreement with the known variability worldwide

The profiles in the database show a wide range of carbon concentrations, taken as a whole or within a same soil type, whatever the depth. Such a variability within the same soil type is in agreement with worldwide surveys. Tifafi et al. (2018) showed that besides the differences between the different worldwide datasets that can be linked

to methodological issues; there remains a large variability that represents natural diversity of soil worldwide. The SoilGrids database (Hengl et al., 2014) (Figure 2.11) highlights a similar variability with a stock ranging from 0 to 560 g/kg and 0 to 500 g/kg for the topsoils of Gleysol and Cambisol respectively. This shows that even if unbalanced, our database captures most of the natural variability of soil, at least from a soil carbon concentration point of view.

Soil type can explain part of the variability of carbon concentration (Figure 2.10). Andosol, for both topsoil and deepsoil, shows the highest concentration of soil carbon. Beyond this observation, there remain some trends between soil types for topsoil carbon content. For instance, Cambisol, Gleysol then Podzol appear to have a higher carbon content than the other types of soil. The pattern for deepsoil carbon content is less clear. Nonetheless, following Andosol, Ferralsol shows a higher carbon content than other soil types. This general relationship between general carbon content in the top layers and in depth with soil type should be captured by modeling. Besides soil type, climate and the associated vegetation can explain part of the variability of carbon concentrations. Gleysols cover 7.2% of the total land surface; they are found in polar regions, the tropics and subtropics, and can be either natural or cultivated. With such a large distribution worldwide, considerable variability is unsurprising. Apart from Gleysols, a distribution trend appears within the soil types with the highest content in Andosol and the lowest in Vertisol (Figure 2.10 upper panel). The high carbon content observed for Andosol is in agreement with the fact that Andosol contains nanominerals of the allophane type that stabilize large amounts of organic matter (Basile-Doelsch et al., 2005). For deep soil, Andosol shows the highest carbon content for the same reason as for topsoil (Figure 2.10, lower panel). Except for Andosol, the carbon content value and variability amongst soil types is much smaller (Figure 2.10, lower panel).

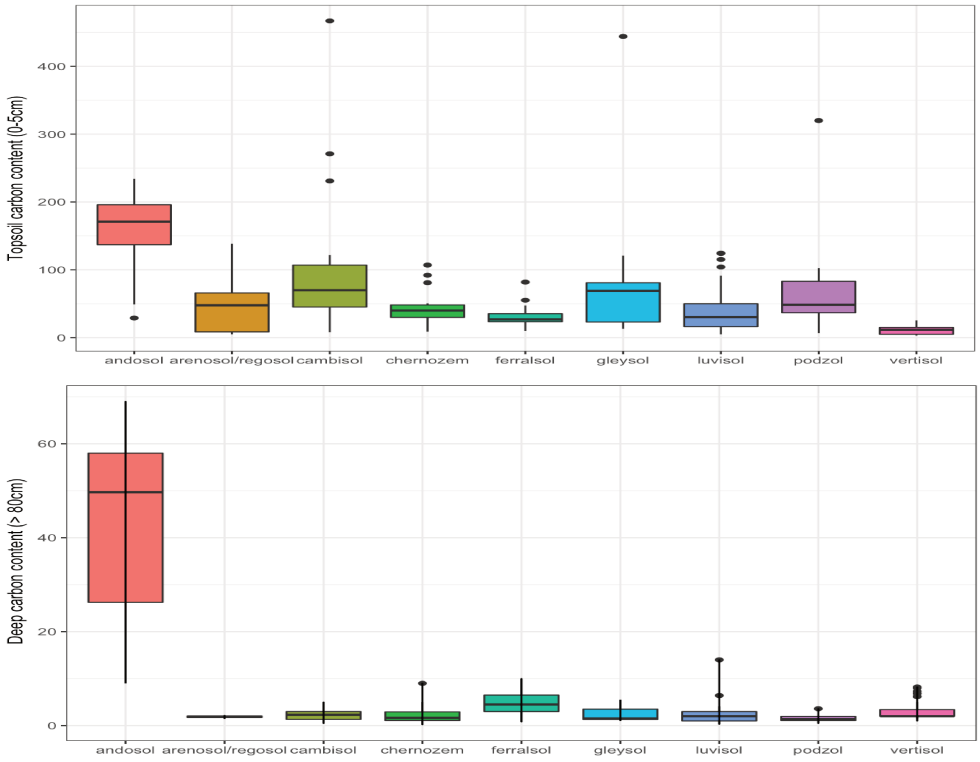


Figure 2.10: The carbon content variations (g/kg) according to the soil type for the top (between 0 and 5 cm) (top panel) and deep soil (greater than 80 cm) layers (bottom panel) obtained on the 125 profiles selected for modeling carbon content dynamics. These Box plots underline quartiles: the extreme of the lower whisker, the lower hinge, the median, the upper hinge and the extreme of the upper whisker.

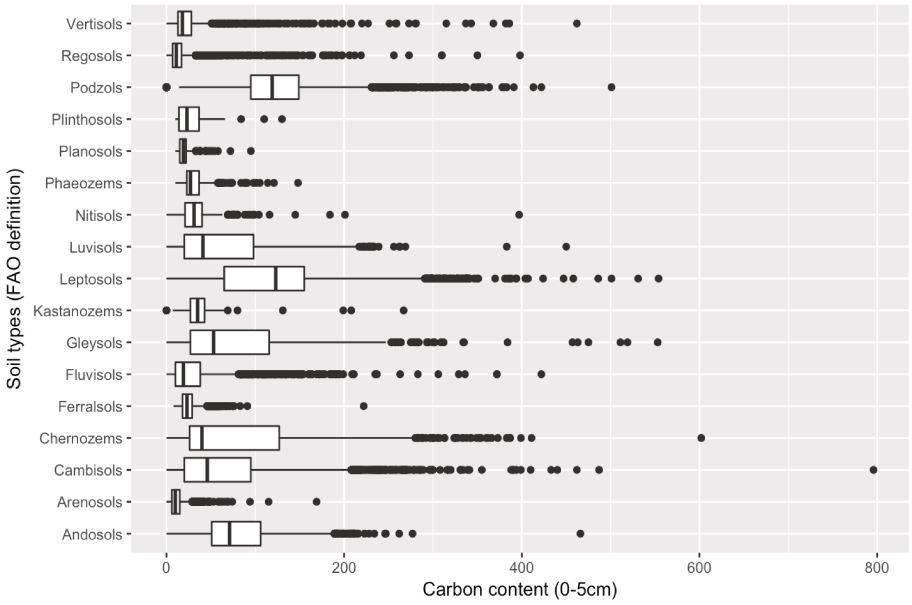


Figure 2.11: The variation in topsoil carbon content (g/kg) recorded among the different soil types (FAO classification) according to the SoilGrids database (http://www.isric.org/explore/soilgrids).

2.2 The take-home messages of the database

As a conclusion regarding the handling of the database:

1. The multicollinearity problem is handled by choosing the five most uncorrelated covariates among the potential 33 numerical climate covariates. The selected climatic factors are: mean annual temperature, mean annual precipitation, seasonal shift between the warmest and coldest months, lowest monthly precipitation and aridity index.

2. The strongly unbalanced experimental design of soil and ecosystem type is improved by merging the soil types and ecosystem types that share similar features according to soil scientists expertise. This has the effect of increasing the number of observations per category and thus increasing the accuracy of the estimators (see Appendix 7.3). A considerable improvement is ensured for the design of the soil types and to a lesser extent for the design of the ecosystem types.

3. The database provides a homogeneous representativeness of intermediate climates, *i.e.* tropical, warm temperate and snow climates, leaving out extreme climates (arid and polar). Ranges of mean annual temperature, mean annual precipitation and seasonal shift are wide and homogeneously swept.

4. The poor distribution of the sampling years for sites before the nuclear bomb period (79% profiles sampled in the 1990s) will not be representative of the range of variation of the $F^{14}C$ for the upper levels as widely as expected in Earth surface.

5. The radiocarbon and carbon content measurements vary between 3 and 73 observations for the sampled sites. The number of observations per profile is mainly from 4 to 10. This is sufficient to adjust the profile shape (4 parameters, see later on). Furthermore, 4 to 10 observations avoids giving too much weight to particular points without losing the specific structure of the profiles.

The topsoil carbon content varies widely according to the soil type. This large variation is in agreement with the worldwide soil carbon database "SoilGrids".

CHAPTER 3

STOCHASTIC SEARCH VARIABLE SELECTION OF MIXED COVARIATES FROM A LATENT LAYER: APPLICATION TO HIERARCHICAL MODELING OF SOIL CARBON DYNAMICS

This chapter introduces a published article on the 13^{th} of September 2018 in the "Journal de la Société Française de Statistique".

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

 ISSN: 2102-6238

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 kg/m^2 for arenosols to more than 10 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

 ISSN: 2102-6238

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

 ISSN: 2102-6238

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

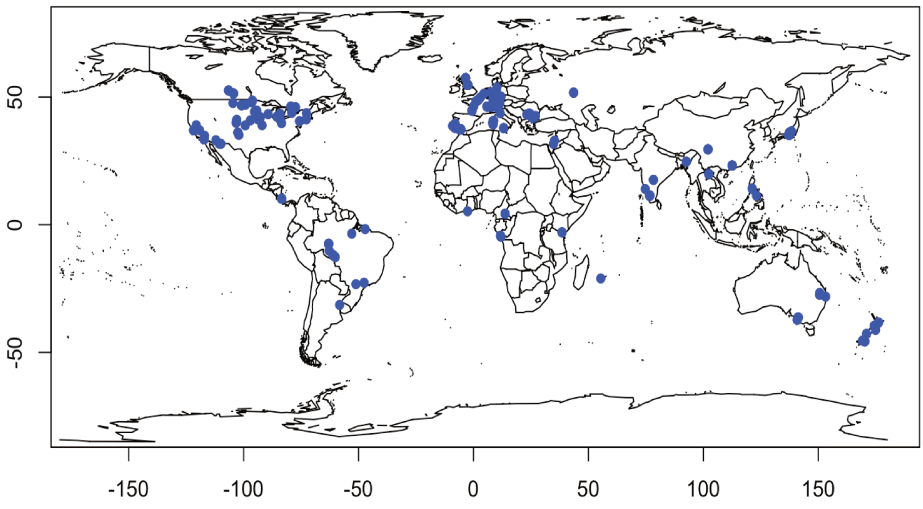


FIGURE 1. *Geographical locations of soil $F^{14}C$ sites.*

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

 ISSN: 2102-6238

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

 ISSN: 2102-6238

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:

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

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

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

 ISSN: 2102-6238

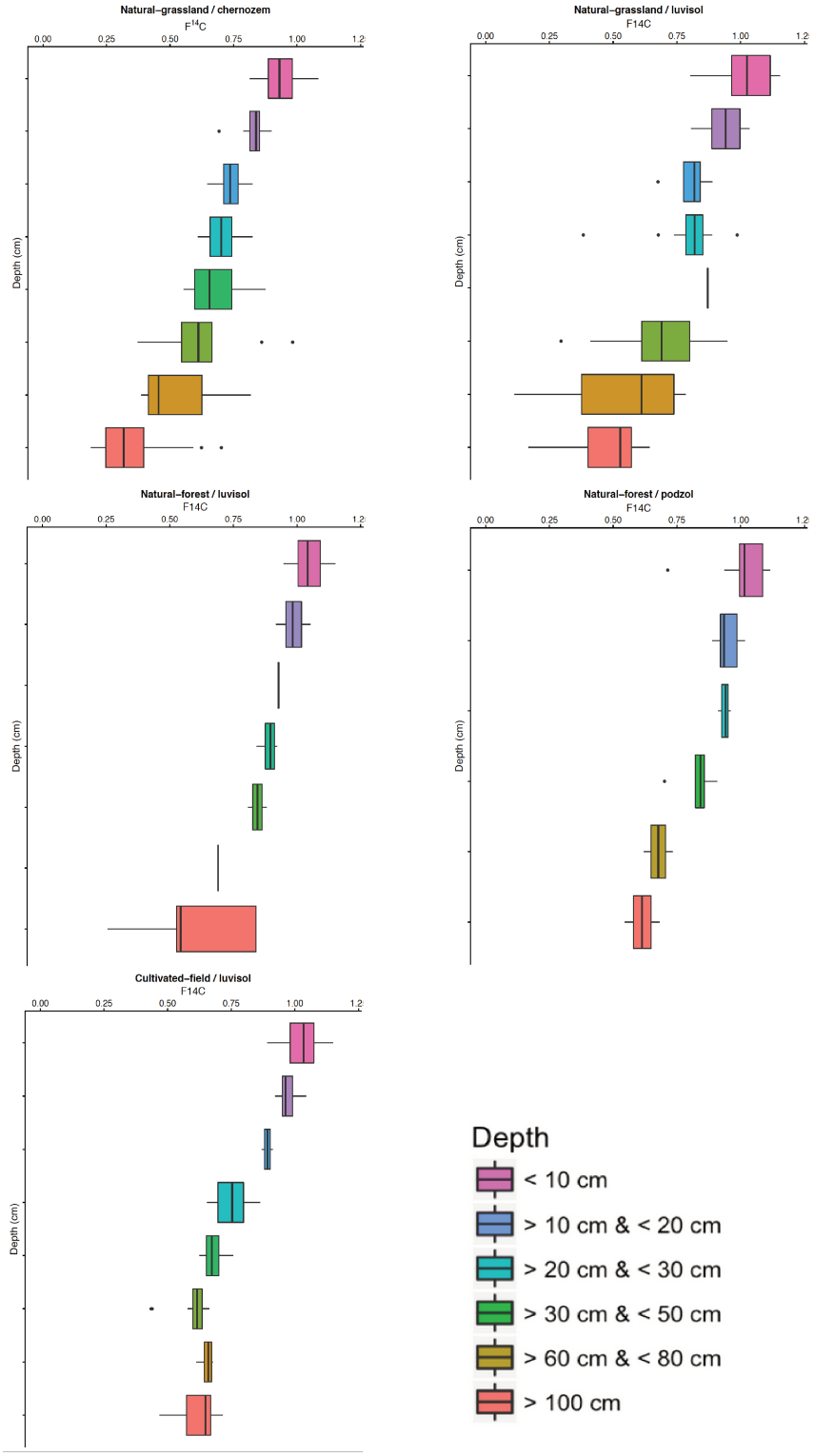


FIGURE 2. *The variation of radiocarbon versus depth is represented on boxplots for the most frequent combinations of ecosystem and soil types. Natural-grassland / chernozem (11 profiles), Natural-grassland / luvisol (11 profiles), Natural-forest / luvisol (9 profiles), Natural-forest / podzol (6 profiles),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

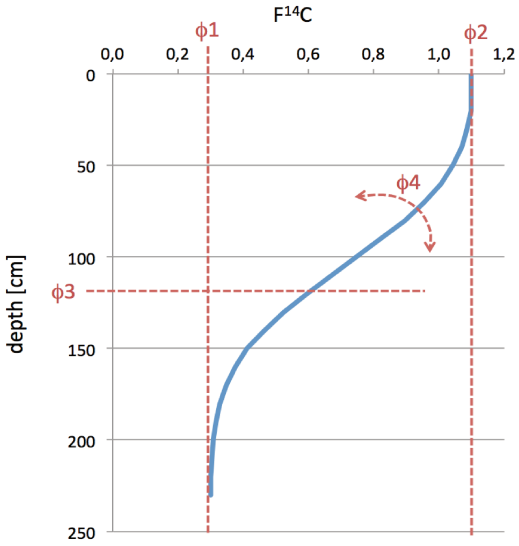


FIGURE 3. *Statistical profile of soil $F^{14}C$ versus depth obtained from Eq. 2.*

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.

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

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

 ISSN: 2102-6238

$\beta_i = (\beta_{i1}, \ldots, \beta_{iP})' \in \mathbb{R}^P$, 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: $\phi_i = (\phi_i(1), \phi_i(2), \ldots \phi_i(S))'$ and $E_i = (E_i(1), E_i(2), \ldots, E_i(S))'$. 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 = $1, \ldots, P$ 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 $j = 1, \ldots, P$. 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

 ISSN: 2102-6238

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 $j = 2, \ldots, K$
    - $\beta_{1j}|I_{1j} \sim (1 - I_{1j}) * N(0, \frac{g_1\sigma_1^2(X'X)^{-1}_{j,j}}{c_1}) + I_{1j} * N(0, g_1\sigma_1^2(X'X)^{-1}_{j,j})$
    - $\beta_{2j}|I_{2j} \sim (1 - I_{2j}) * N(0, \frac{g_2\sigma_2^2(X'X)^{-1}_{j,j}}{c_2}) + I_{2j} * N(0, g_2\sigma_2^2(X'X)^{-1}_{j,j})$
    - $\beta_{3j}|I_{3j} \sim (1 - I_{3j}) * N(0, \frac{g_3\sigma_3^2(X'X)^{-1}_{j,j}}{c_3}) + I_{3j} * N(0, g_3\sigma_3^2(X'X)^{-1}_{j,j})$
    - $\beta_{4j}|I_{4j} \sim (1 - I_{4j}) * N(0, \frac{g_4\sigma_4^2(X'X)^{-1}_{j,j}}{c_4}) + I_{4j} * N(0, g_4\sigma_4^2(X'X)^{-1}_{j,j})$

 ISSN: 2102-6238

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

  For $j = k+1, \ldots, P$ 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.

 ISSN: 2102-6238

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, $X_1,\ldots,X_6$ *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:

$$X_5 = 2\times X_3$$
$$X_6 = X_2 + 1.5\times Z, \quad Z\sim N(0,1)$$

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 $(X_1,\ldots,X_6)$ 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 $\beta_4=(\mathbf{0},\mathbf{0},1,0,0,1,0.8,0.8)$. $\mathbf{0}$ and $\mathbf{1}$ 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 β_1,β_2,β_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:

 ISSN: 2102-6238

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.

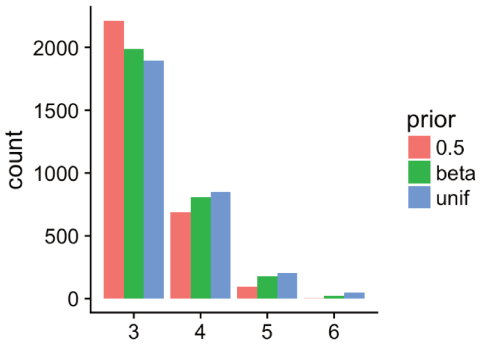


FIGURE 4. *The number of selected covariates identified among the MCMC iterations (nb of iterations = 3000) for the third latent variables (ϕ_3) for the three proposed priors for the probability selection p.*

 ISSN: 2102-6238

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.*

a)

c = 10		c = 100		c = 1000		c = 5000	
Prob	F.C	Prob	F.C	Prob	F.C	Prob	F.C
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

b)

c = 10		c = 100		c = 1000		c = 5000	
Prob	F.C	Prob	F.C	Prob	F.C	Prob	F.C
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)

c = 10		c = 100		c = 1000		c = 5000	
Prob	F.C	Prob	F.C	Prob	F.C	Prob	F.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

d)

c = 10		c = 100		c = 1000		c = 5000	
Prob	F.C	Prob	F.C	Prob	F.C	Prob	F.C
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

 ISSN: 2102-6238

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

 ISSN: 2102-6238

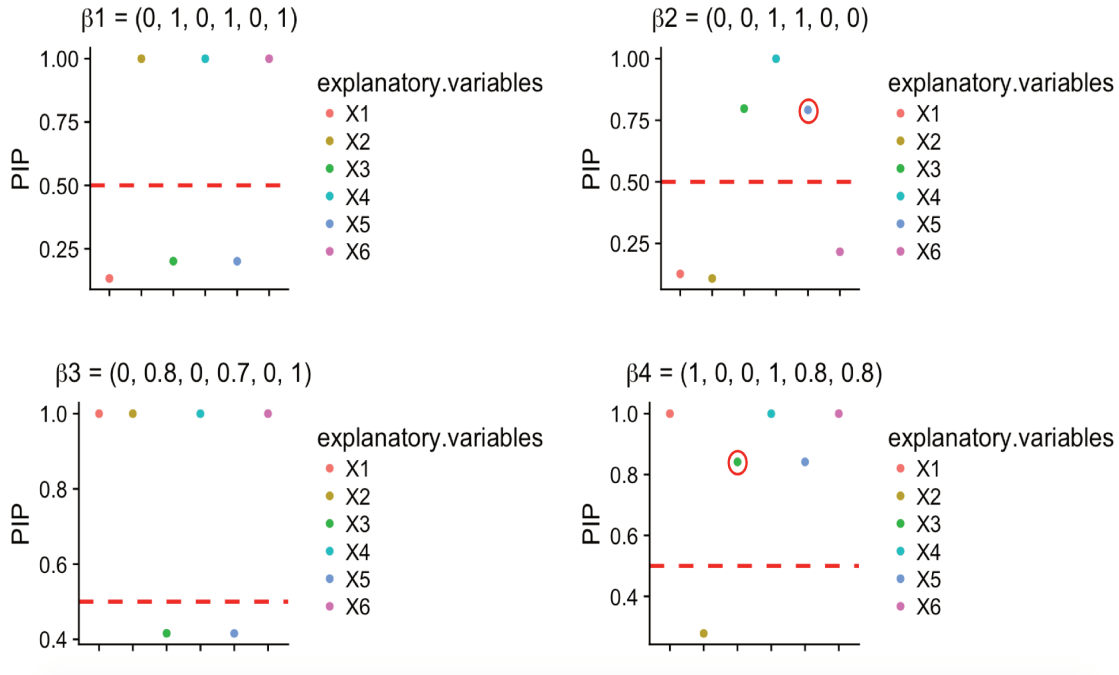


FIGURE 5. ***SSVS*** *evaluation for artificial data including both independent and correlated quantitative covariates. Panels a, b, c, d give the results obtained for the Posterior Inclusion Probability (PIP) separately for each covariate throughout the sub-models identified by the MCMC iterations for 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

 ISSN: 2102-6238

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.

 ISSN: 2102-6238

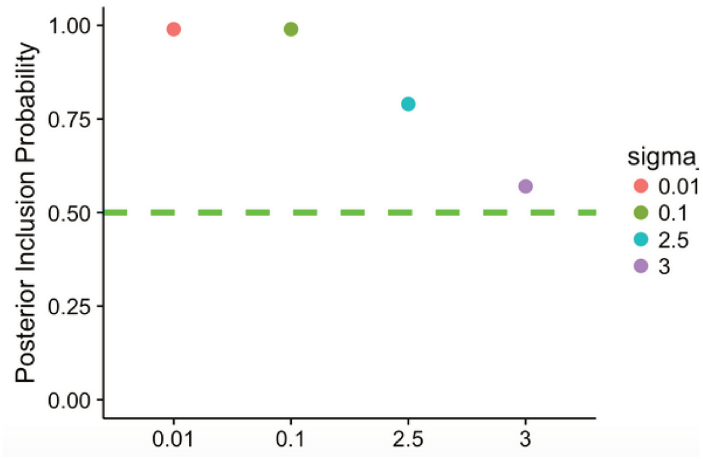


FIGURE 6. *Posterior inclusion probability in relation with σ_1 value. The illustration here is for X_2 included as a descriptor of the ϕ_1 latent layer model, ϕ_i for $i = 2,3,4$ being fixed. The posterior inclusion probability decreases versus increasing values of σ_1. The green dashed line represents 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 $c_i (i = 1,\ldots,4)$ 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.

 ISSN: 2102-6238

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

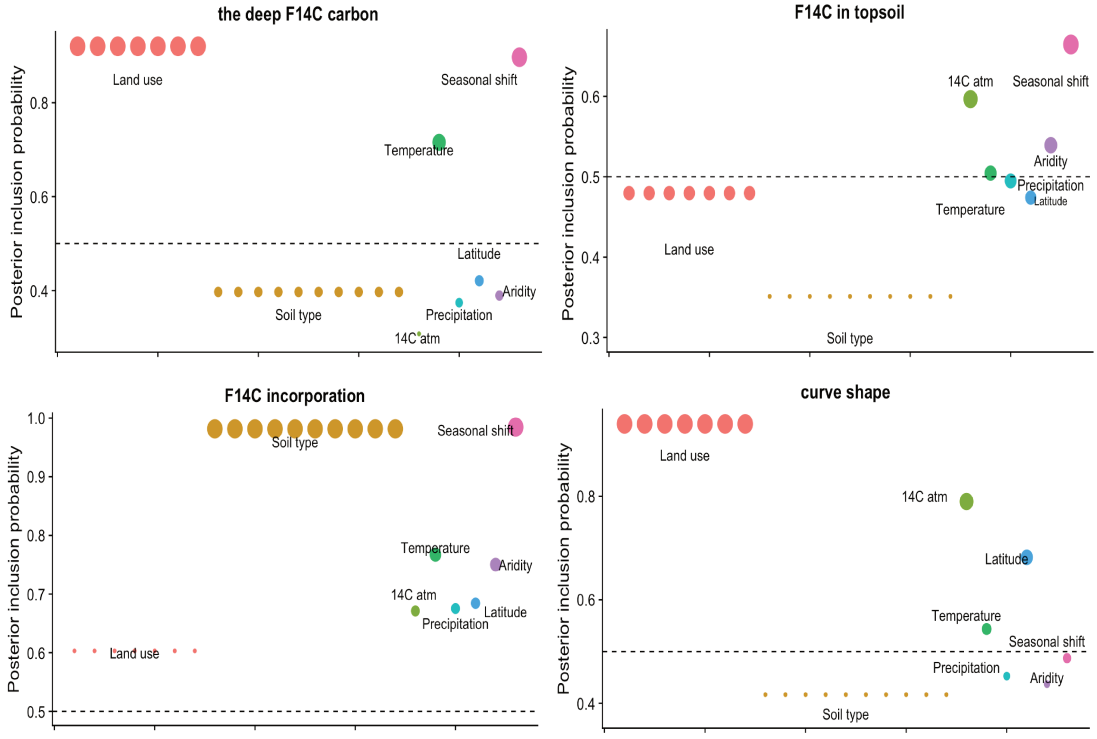


FIGURE 7. *Posterior inclusion probabilities for all explanatory covariates obtained by applying the **SSVS** to the entire real database. The size of points depends on the importance of the posterior inclusion probability.*

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

 ISSN: 2102-6238

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.

 ISSN: 2102-6238

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.

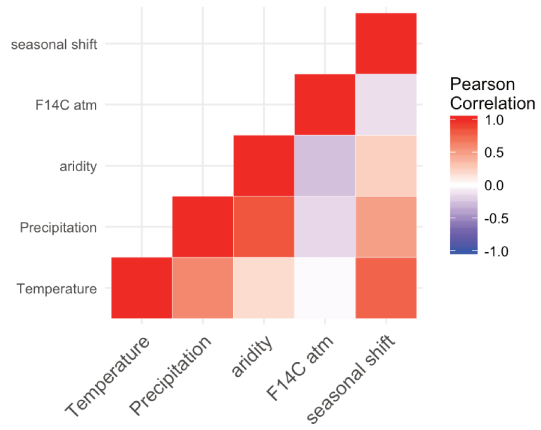


FIGURE 8. *Correlation matrix of the six quantitative explanatory covariates. The darker the color (red or blue), the stronger the correlation between the variables (positive or negative)*

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($y^{rep},\theta$), 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)

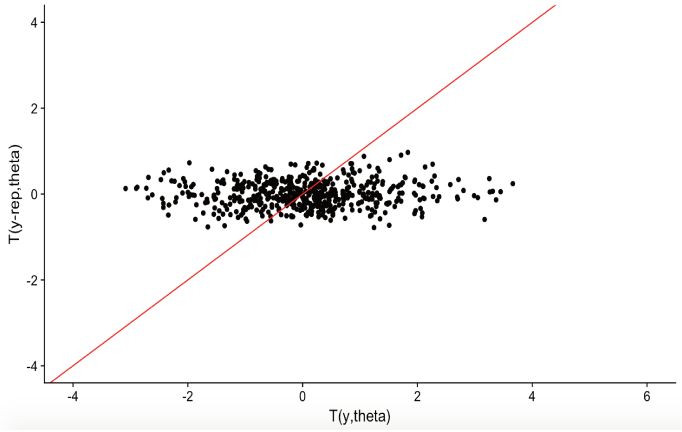


FIGURE 9. *The discrepancy measures $T(X,y^{rep},\theta)$ calculated on replicated data and parameters model θ versus $T(X,y,\theta)$ calculate on real data and θ. The estimated Bayesian p-value is equal to 0.47.*

 ISSN: 2102-6238

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

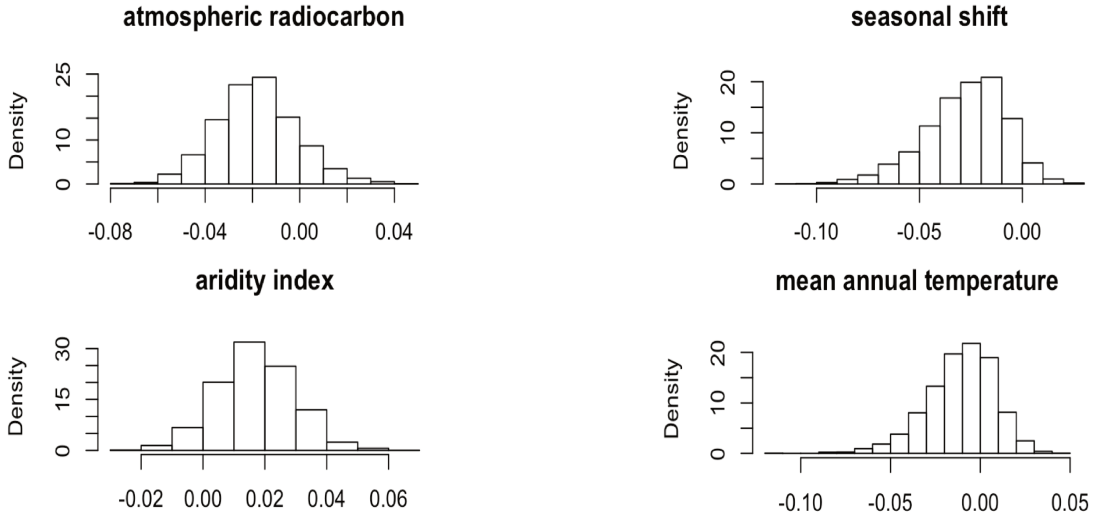


FIGURE 11. *The posterior distribution of the regression effects corresponding to the significant numerical covariates for topsoil radiocarbon latent model (ϕ_2): atmospheric radiocarbon, seasonal shift, aridity index and mean annual temperature.*

 ISSN: 2102-6238

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

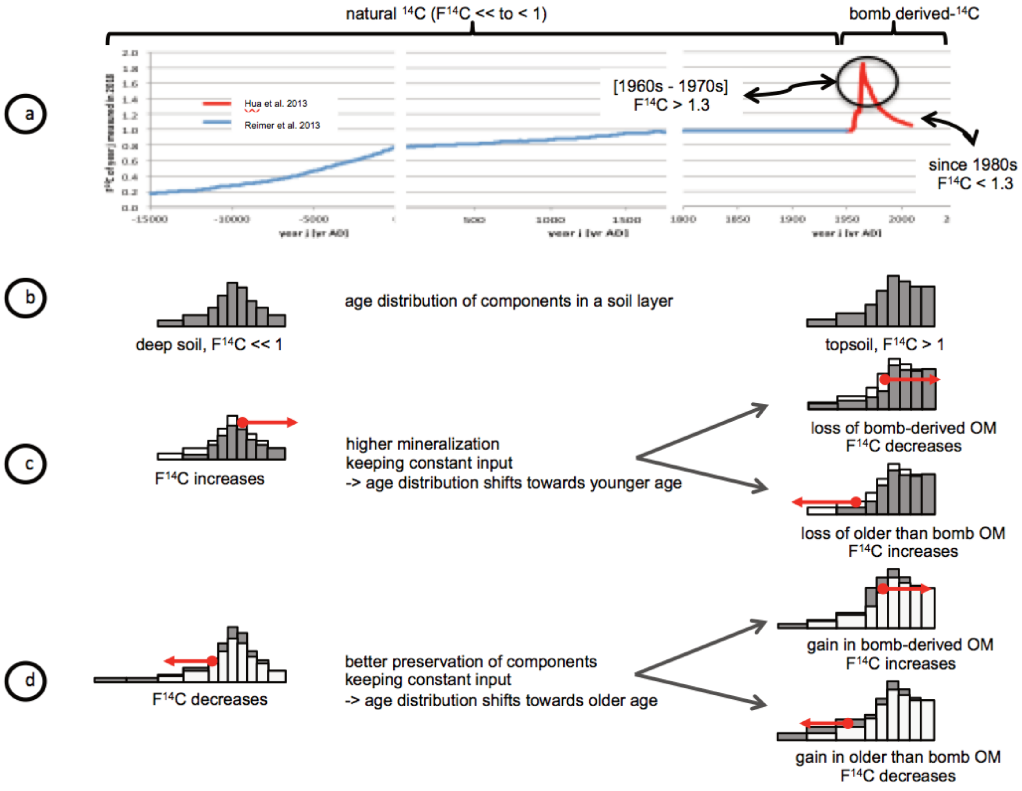


FIGURE 12. *The first graph a shows the variation of the atmospheric $F^{14}C$ concentration over time. The soil was affected specially by above-ground nuclear testing from about 1950 until 1963. Panel b highlights the variation of radiocarbon amount between deep and topsoil. The last 2 panels c and 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.*

 ISSN: 2102-6238

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.

 ISSN: 2102-6238

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

 ISSN: 2102-6238

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*.

 ISSN: 2102-6238

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*.

 ISSN: 2102-6238

Note: In figure 5 of the article, the vector β_3 at the top of the third panel should be $\beta_3 = (1, 0.8, 0, 0.7, 0, 1)$ instead of $\beta_3 = (0, 0.8, 0, 0.7, 0, 1)$

Rectified references:

- 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.
- Martin, W., Smith3, M., Jolivet, B., and Arrouays (2011). Spatial distribution of soil organic carbons stocks in France.

CHAPTER 4

BAYESIAN SELECTION APPROACHES FOR CATEGORICAL PREDICTORS USING JAGS

This chapter introduces a submitted paper, it is for this reason that it will not be available in this manuscript.

CHAPTER 5

APPLICATIONS TO THE SOIL CARBON DATABASE

In Chapter 3, the performance of the Stochastic Search Variable Selection (SSVS), originally built to select the significant numerical predictors, was adapted to the hierarchical non linear model with latent variables proposed for modeling soil radiocarbon dynamics. The SSVS had to be modified to handle the presence of categorical predictors such as the soil type and the ecosystem type of the sampled profile. In addition, the results obtained on artificial data generated according to the proposed statistical model highlighted that the detection of some significant categorical predictors can be misleading with SSVS. Further investigation done on the real data indicated that the prediction and model fitting were better after the inclusion of some categorical predictors detected as non significant by the SSVS. The results of the SSVS on the soil radiocarbon dynamics, published in the SFDS journal (see Chapter 3), gave rise to new questions: How can Bayesian selection methods handle the presence of categorical predictors? What are the Bayesian selection methods appropriate for categorical predictors that already exist in the literature? Which Bayesian selection methods to choose?
These questions led to the submission of the second article "Bayesian selection approaches for categorical predictors using JAGS" from which Chapter 4 derived. Three Bayesian Selection approaches appropriate for categorical predictors: Bayesian Group Lasso with Spike and Slab, Bayesian Sparse Group Selection and Bayesian Effect Fusion were tested on a simple linear model with categorical predictors in order to highlight the prior specifications for each Bayesian Selection method rather than the complexity of the statistical model.

In this chapter, we now test the performance of the Stochastic Search Variable Selection, introduced in Chapter 3, for the numerical predictors (mean annual temperature, aridity index, etc.) and the three Bayesian selection approaches explored in Chapter 4 for the categorical predictors (soil type and ecosystem type) on the latent layers within the framework of a non linear hierarchical model. Here, it should be pointed out that the framework of the application becomes more complex.

This chapter will be organized as follows: Section 5.1 recaps the structure of the statistical model proposed for the soil radiocarbon dynamics and the climatic and environmental factors considered for this statistical analysis. It also introduces the structure of the statistical model for soil carbon content modeling and the potential climatic and environmental factors. Section 5.2 will explore the Bayesian Full model, in which all the explanatory predictors are included, and the implementation of Bayesian Selection approaches previously introduced in the manuscript. This section will be divided into two parts: the first part applies these methods to the soil radiocarbon dynamics and the second one will address the application of these methods to the soil carbon content dynamics. Following this outline, all the Bayesian methods are then compared with respect to both soil radiocarbon and soil carbon content dynamics based on the Bayesian selection criteria for model comparisons.

After choosing the best subset of climatic and environmental factors for soil radiocarbon and soil carbon content dynamics, Section 5.3 will touch on the physical interpretation of the selected climatic and environmental factors. This section will also comment on the expectations of soil scientists versus the selected and the non selected climatic and environmental predictors.

Finally, Section 5.5 will underline the predictive model applications in a context of global climatic and land use changes.

5.1 Recap of soil carbon database and statistical modeling

5.1.1 Recap of the statistical model and the potential explanatory predictors for soil radiocarbon dynamics

A - The structure of the statistical model for soil radiocarbon response

As presented in Chapter 3, for each site $s \in \{1 : S\}$ and each depth $x \in \{1 : m_s\}$, where m_s is the number of measurements available for site s, the $F^{14}C$ content experimental record is modeled by:

$$F^{14}C(s,x) = \phi_1(s) + (\phi_2(s) - \phi_1(s)) * exp\left(-\frac{x}{\phi_3(s)}\right)^{\phi_4(s)} + \varepsilon(s,x) \quad \varepsilon(s,x) \sim N(0,\sigma^2) \tag{5.1}$$

- ϕ_1: deep radiocarbon;
- ϕ_2: topsoil radiocarbon;
- ϕ_3: is related to the inflection point of the curve;
- ϕ_4: describes the more or less rapid decrease of $F^{14}C$;

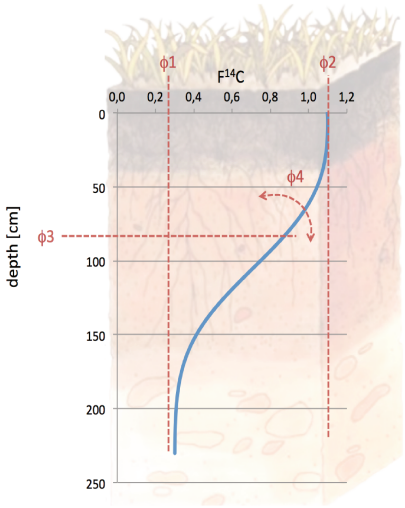


Figure 5.1: Statistical profile of soil $F^{14}C$ versus depth obtained from Equation 5.1.

B - The potential climatic and environmental predictors for soil radiocarbon dynamics

In Chapter 3, the latitude was considered as a potential explanatory predictor for soil radiocarbon dynamics. Latitude was initially included in the statistical model, since the soil carbon varies widely across different latitudes. However, on further consideration, we decided to remove latitude from the statistical model, since the variation in soil radiocarbon is not directly affected by the latitude beyond the relationship (correlation) of the latitude with the temperature.
Thus, the six climatic factors listed in Figure 5.2 were chosen in order to reduce the multicollinearity problem (section 2.1.3.1 of Chapter 2).

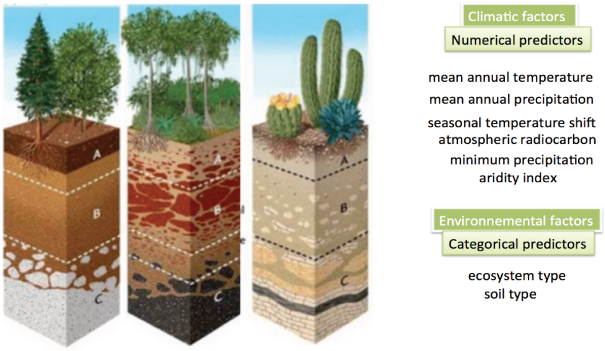


Figure 5.2: The climatic and environmental factors that potentially impact soil radiocarbon dynamics.

As stated in Chapter 3, the integration of climatic and environmental factors is done by considering a hierarchical model with latent layers. A linear model is used to link the potential explanatory predictors and the latent layers (Equation 5.2):

$$\phi_i = X\beta_i + \varepsilon_i \quad \varepsilon_i \sim N_S(\mathbf{0}, \sigma_i^2 * I) \quad i = 1,2,3,4 \tag{5.2}$$

where, $\phi_i = (\phi_i(1), \ldots, \phi_i(s), \ldots, \phi_i(S))$, $\beta_i \in \mathbb{R}^P$, $\varepsilon_i = (\varepsilon_i(1), \ldots, \varepsilon_i(s), \ldots, \varepsilon_i(S))$ $\mathbf{0}$ is the vector of 0 components and I $\in M_{S,S}(\mathbb{R})$ is the identity matrix.

To handle the presence of categorical predictors in the latent linear models, a treatment contrast is used to build the design matrix $X \in M_{S,P}(\mathbb{R})$ (S = 131, P = 20).
Here, six numerical predictors and two categorical predictors with six and nine levels respectively are included in the latent layer models. The category having the largest number of observations is chosen as the baseline level. Thus, "natural-forest" (37% of the total number of profiles) and "Luvisol" (27% of the total number of profiles) were chosen as the baseline levels respectively for the ecosystem and soil type (see Figure 2.4 in Chapter 2).
The abbreviations of the potential climatic and environmental factors are recalled in Table 5.1.

Potential covariates	Abbreviation	Potential covariates	Abbreviation
ecosystem type	Land	soil type	Soil
mean annual precipitation	MAP	aridity index	AI
mean annual temperature	MAT	seasonal temperature shift	Dif_T
minimum precipitation	min_P	atmospheric $F^{14}C$ of the sampling year	$F^{14}C$atm

Table 5.1: Abbreviations of the climatic and environmental predictors considered for soil radiocarbon modeling.

5.1.2 The structure of the statistical model for soil carbon content dynamics

According to the observed profiles of soil carbon dynamics (Figure 5.3), we proposed for each site $s \in \{1 : S\}$ and each depth $z \in \{1 : m_s\}$. The structure of the statical model for soil carbon content dynamics is given by:

$$C(s,x) = \omega_1(s) + (\omega_2(s) - \omega_1(s)) * exp\left(-\frac{x}{\omega_3(s)}\right) + \varepsilon \quad \varepsilon \sim N(0, \sigma_c^2) \tag{5.3}$$

$$\omega_i = X^*\eta + \varepsilon_i^* \quad \varepsilon_i^* \sim N(0, sd_i^{2*}) \quad \text{for} \quad i = 1,2,3 \tag{5.4}$$

where, for i = 1, 2, 3, $\omega_i \in \mathbb{R}^S$ and $\varepsilon_i^* \in \mathbb{R}^S$ $X^* \in M_{S,P'}(\mathbb{R})$, with P' = 19, is the design matrix constructed using treatment contrast. The climatic and environmental predictors considered for modeling the soil carbon content dynamics are recalled in Table 5.2:

Potential covariates	Abbreviation	Potential covariates	Abbreviation
ecosystem type	Land	soil type	Soil
mean annual precipitation	MAP	aridity index	AI
mean annual temperature	MAT	seasonal temperature shift	Dif_T
minimum precipitation	min_P		

Table 5.2: Abbreviation of the eight climatic and environmental predictors considered for soil carbon content modeling.

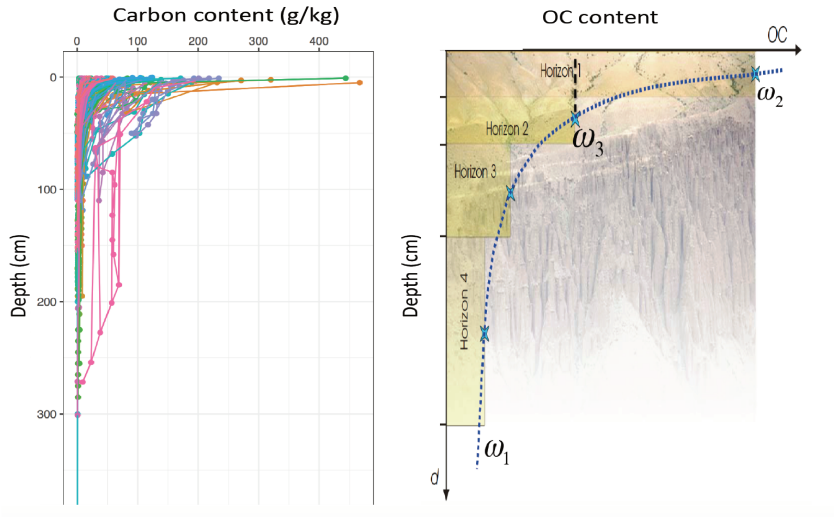


Figure 5.3: The left panel shows the real variation of the carbon content with depth for all studied sites. The right panel underlines the structure of the statistical model proposed according to the left panel. ω_1 defines the deep soil carbon content, ω_2 represents the topsoil soil carbon content and ω_3 is related to the point from which the curve changes decay rate.

5.1.3 Depth modeling for both radiocarbon and soil carbon content vertical dynamics

For each profile in the database, the measurements of radiocarbon and soil carbon content were taken from soil layers characterized by upper and lower levels. The question was which depth level to take for the statistical modeling: upper, lower, mean or median levels?

For our statistical analysis, we chose to use the mean depth level of the corresponding soil layer. In fact, for the deep layers, it is preferable to measure radiocarbon or soil carbon content at a depth tending towards the lower level since the deep layers are characterized by a low value of $F^{14}C$ (before the nuclear bomb peak) and soil carbon content. In contrast, for the topsoil layers, we do not know exactly which depth level the measurement taken corresponds to. In fact, the topsoil carbon is much more sensitive to external factors such as the Net Primary Production (NPP) and the climatic conditions than the deep layers. In addition, a soil layer integrates more ^{14}C for the levels corresponding to the nuclear bomb peak than the levels corresponding to the period after the peak. Some soil layers will also be characterized by an upper level rich in soil carbon content and poor in radiocarbon which is the opposite of the behavior observed at the lower level. For these reasons, we chose the mean level as a good compromise to represent the measurements of radiocarbon and carbon content.

In Figure 5.4, we display a $F^{14}C$ profile sampled in 1986 by Becker-Heidmann et al. (2002) under a cultivated field and a Vertisol soil type. This site was characterized by a 2cm-step $F^{14}C$ measurement to 36cm depth, then a 4cm-step (first panel). In order to study the impact of level thickness on the statistical modeling, we averaged the original data to double the sampling step, *i.e.* 8 and 12 cm (second panel) to reach a 48cm-step sampling (third panel). The blue curve represents the fitted statistical model proposed for soil radiocarbon dynamics, based on an

optimization algorithm.

For example, when increasing the level thickness from 2 cm to 4cm and 48cm, the topsoil and deep estimated $F^{14}C$ values, given by the blue curve, remain almost identical. On the other hand, when the level thickness strongly increases from 2cm to 48cm, the curve shape will miss some features such as the increase in the soil radiocarbon observed after the first measures collected (Figure 5.4, first and second panels).

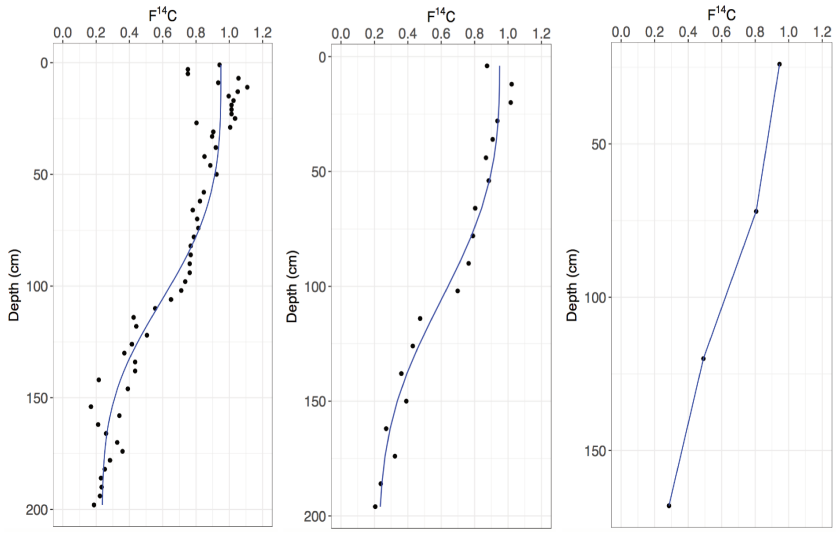


Figure 5.4: Comparison of the same $F^{14}C$ profile with different values of layer thickness (2cm-step $F^{14}C$ measurement to 36 cm depth, then a 4 cm-step (first panel), 8 and 12 cm (second panel) and 48 cm-step sampling(third panel)) where the sample was collected in 1986 by Becker-Heidmann et al. (2002) under a cultivated field (Vertisol).

5.2 Bayesian modeling and Bayesian selection methods

5.2.1 Application to soil radiocarbon dynamics

5.2.1.1 Full Bayesian model

5.2.1.1.1 Full Bayesian model specification and constraints

After thorough deliberation on the statistical Bayesian model proposed in Chapter 3, the following constraints were added to the Bayesian model likelihood and latent variables:

The constraints of the Full Bayesian model

1. The unit $F^{14}C$ used to express the soil radiocarbon always has a positive value. To ensure the generation of positive values for the soil radiocarbon response, two solutions can be tailored:
   - ✓ Assuming a Normal distribution for the logarithm transformation of the response $F^{14}C$. The logarithm transformation can be applied since the unit $F^{14}C$ has a positive value and cannot be equal to 0.
   - ✓ Considering a Normal distribution truncated at 0 to model soil radiocarbon.
2. According to the interpretations of ϕ_1, ϕ_2, ϕ_3 and ϕ_4, these four latent variables should always be positive:
   (a) Assuming a Normal distribution truncated at zero for the first three latent variables ϕ_1, ϕ_2 and ϕ_3. These three latent variables have a physical interpretation. The truncated Normal was favored in order to interpret directly the effect of climatic and environmental factors on these latent variables.

(b) A logarithm transformation is used to model the fourth latent variable, ϕ_4.

For the Bayesian inference, with no prior information about the precision of the parameters, the scale and the hyperparameter shape of the Gamma distribution must be chosen so as to give a very disperse prior. The most widely used parametrization of the Gamma distribution is to have the same number in both hyperparameter positions. Thus, by choosing 0.001 for both hyperparameters, the precision has a mean 1 and a large variance of 1000.
For the regression effects for latent linear models, we assumed Zellner's g priors. In fact, Zellner's g prior (Zellner, 1986) is based on the idea that the regression effect estimation should be invariant to changes in the scale of the regressors. Some linear algebra shows that this condition is satisfied if the mean and the variance of the Normal distribution prior on the regression effect β are equal to 0 and $k(X'X)^{-1}$ respectively. A popular specification is to set $k = g\sigma^2$ for positive values of g. The choice of g can be based on many popular model selection criteria, such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and others. However, assuming a prior on g has the advantage of avoiding paradoxes such as "Bartlett's Paradox" and the "Information Paradox". Briefly, "Bartlett's Paradox" states that the null model would always be preferred to any other model when g$\to \infty$. On the other hand, when the coefficient of determination $R^2 \to 1$, the Bayes factor converges to a constant instead of going to ∞ as we consider that the datasets fit the model better. This problem is called the "Information Paradox". For the Full Bayesian model, we consider a vague Uniform prior on g, assigning the same weight to all possible values of g.

FULL BAYESIAN MODEL FOR SOIL RADIOCARBON MODELING

- **Likelihood**:

  for each site $s \in \{1 : S\}$ (S = 131) and each depth $x \in \{1 : m_s\}$, the likelihood is written as:

  – $F^{14}C(s,x) \sim N_+\left(g(\phi(s),x), \sigma^2\right)$ (truncated Normal distribution)

  or

  – $log\left(F^{14}C(s,x)\right) \sim N\left(log(g(\phi(s),x)), \sigma^2\right)$ (logarithm transformation)

  where, $g(\phi(s),x) = \phi_1(s) + (\phi_2(s) - \phi_1(s)) * exp\left(-\frac{x}{\phi_3(s)}\right)^{\phi_4(s)}$

- **Latent variables:**
  for each site s, the linear latent models are defined as:

$$\phi_i(s) \sim N_+\left(X[s,] * \beta_i, \sigma_i^2\right) \quad i = 1,2,3$$
$$log(\phi_4(s)) \sim N\left(X[s,] * \beta_4, \sigma_4^2\right)$$

- **Priors:**

  – $1/\sigma^2 \sim Gamma(0.001, 0.001)$
  for i = 1, 2, 3, 4, P = 20 and S = 131:

  – $1/\sigma_i^2 \sim Gamma(0.001, 0.001)$

  – $\beta_i \sim N(0, g_i\sigma_i^2(X'X)^{-1})$ where $\mathbf{0} \in \mathbb{R}^P$ and $X \in M_{S,P}(\mathbb{R})$

  – $g_i \sim Uniform(5, 5000)$

Compared to the soil radiocarbon statistical model published in the SFDS journal, three changes were introduced: 1- the statistical analysis is applied on a new version of the database which contains 131 profiles instead of 104 (27 profiles have been added to the database of Chapter 3) 2- the positivity constraint has been added to ensure the generation of positive values for the responses $F^{14}C$, 3- the latitude is no longer considered as a significant predictor since it is linked to the temperature and does not influence the soil radiocarbon response directly, 4- Normal distributions truncated at zero are proposed for the first three latent variables (ϕ_1, ϕ_2 and ϕ_3) in order to ensure the generation of positive values and to have a direct physical interpretation of the regression effects of climatic and environmental predictors.

5.2.1.1.2 Results of the Full Bayesian model

All models are wrong, but some are useful (Box, 1976). Bayesian model checking and model selection criteria cannot tell us which model is true, but can tell us how well each model fits the data. One can distinguish among the following criteria: the Deviance Information Criterion (DIC), the p-value of the Posterior Predictive Check (P.P.C), the error on the validation sets of k-fold Cross Validation (C.V) and the Coverage of Bayesian credible intervals on validation sets (more details are given in Appendix 7.2.1).
The full Bayesian model was tested under two scenarios: 1- a Normal distribution truncated at 0 was used to model the $F^{14}C$ response (additive model), 2- a log transformation was applied to the $F^{14}C$ response (multiplicative model). Before testing the Full Bayesian model under these two scenarios, a preliminary study was done. We estimated the latent variables ϕ_1, ϕ_2, ϕ_3 and ϕ_4 of the non linear mean structure by minimizing the square error between the real observations and the replicates generated according to the statistical model. For each site and each depth, the residuals between the $F^{14}C$ observations and the estimated non linear means are plotted in Figure 5.5.

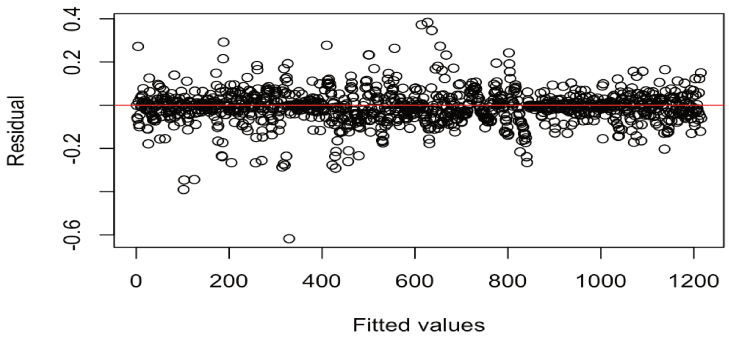


Figure 5.5: Residuals versus fitted values for all $F^{14}C$ profiles and all depth measurements. Positive values for the residual mean that the prediction is too low, and negative values mean that the prediction is too high; 0 means that the estimation was exactly correct.

Figure 5.5 shows that the variance of the residuals is homogeneous throughout the dataset. This result favors the scenario of considering an additive model with a truncated Normal distribution on the radiocarbon observations. After testing the Full model under both of these scenarios, this was confirmed and a better data fitting is observed under an additive model (Table 5.3).

Models	DIC	p-value (P.P.C)
additive model (truncated Normal distribution)	-2324	0.58
multiplicative model (log transformation)	-1371	0.71

Table 5.3: Deviance Information Criterion (DIC) comparison between the Full additive model and the multiplicative Full model. The model with the lowest DIC is preferred to models with larger DIC.

The test quantity $T(y,\theta)$ used to compute the p-value of the Posterior Predictive Check (P.P.C) is equal to the mean of the difference between the real radiocarbon observation and the non linear mean obtained according to the latent variables $\hat{\phi}_1, \hat{\phi}_2, \hat{\phi}_3$ and $\hat{\phi}_4$ sampled from their posterior distributions. A better agreement between the model and the dataset was achieved under the additive model (0.58 p-value closer to 0.5 and farther from 1).

Based on the Deviance Information Criterion and the p-value of the Posterior Predictive Check criterion, an additive model with a truncated Normal distribution on $F^{14}C$ was chosen for soil radiocarbon modeling.

5.2.1.2 Bayesian Group Lasso with Spike and Slab prior

5.2.1.2.1 BGL-SS model specification and choice of hyperparameters

As stated in Chapter 4, the Bayesian Group Lasso with Spike and Slab prior (BGL-SS) is the simplest Bayesian selection approach used for both categorical and numerical predictors. Furthermore, this Bayesian selection method requires few hyperparameters to tune. Within the framework of the non linear hierarchical model, the BGL-SS was been applied to the latent linear models for ϕ_1, ϕ_2, ϕ_3 and ϕ_4. Thus, the BGL-SS model is specified as follows:

BAYESIAN GROUP LASSO WITH SPIKE AND SLAB PRIOR

- **Likelihood**:
  for each site $s \in \{1:S\}$ (S = 131) and each depth $x \in \{1:m_s\}$, the likelihood is written as:
  - $F^{14}C(s,x) \sim N_+\left(g(\phi(s),x), \sigma^2\right)$

  where, $g(\phi(s),x) = \phi_1(s) + (\phi_2(s) - \phi_1(s)) * exp\left(-\frac{x}{\phi_3(s)}\right)^{\phi_4(s)}$

- **Latent variables:**

  for each site s, the linear models are defined us:

$$\phi_i(s) \sim N_+\left(X[s,] * \beta_i, \sigma_i^2\right) \quad i = 1,2,3$$
$$log(\phi_4(s)) \sim N\left(X[s,] * \beta_4, \sigma_4^2\right)$$

- **Priors**:
  - $1/\sigma^2 \sim Gamma(0.001, 0.001)$
  - for the fourth latent variable ϕ_4, we propose a vague Normal prior:
    $\beta_4 \sim N_P(\mathbf{0}, 100 * \mathbf{I})$ where $\mathbf{0} \in \mathbb{R}^P$, $\mathbf{I} \in M_{P,P}(\mathbb{R})$ and P = 20
  - Intercepts:

    $\beta_{0,i} \sim N(0,1)$ for i = 1, 2
    $\beta_{0,3} \sim N(0, 1000)$
  - for i = 1, 2, 3, 4:
    $1/\sigma_i^2 \sim Gamma(0.001, 0.001)$

- **Priors**:
  - Here, the eight potential predictors are divided into six numerical covariates and two categorical ones. m_g is defined as follows:

$$m_g = \begin{cases} C_g & \text{the number of levels (without the baseline) for categorical predictors} \\ 1 & \text{for the numerical covariates} \end{cases}$$

  - for g = 1,. . . , 8 and i = 1, 2, 3:

$$\beta_{i,g} \sim (1-\pi_{i,g}) N_{m_g}(0, \sigma_i^2 \tau_{i,g}^2 I_{m_g}) + \pi_{i,g} \delta_0(\beta_{i,g})$$

$$\tau_{i,g}^2 \sim Gamma(\tfrac{m_g+1}{2}, \tfrac{\lambda_i^2}{2})$$

$$\lambda_i^2 \sim Gamma(0.001, 0.001)$$

$$\pi_{i,g} \sim Bernoulli(p_{i,g})$$

$$p_{i,g} \sim Beta(2,2)$$

Choice of hyperparameters

- **Shrinkage parameter (λ):** Bayesian inference has the advantage of considering the shrinkage coefficient as a parameter and of suggesting a prior distribution. Bayesian inference helps us to save time since we do not have to apply a cross validation to choose, among several proposed values, the best shrinkage parameter value. We assume a vague Gamma prior of a mean equal to 1 and a variance of 1000.
- **Prior inclusion probability**: Chapter 4 showed, on a simulated study, that the BGL-SS is not sensitive to the prior inclusion probability. Thus, we considered a Beta distribution with both hyperparameters equal to 2 on the prior inclusion probability (Figure 5.6):

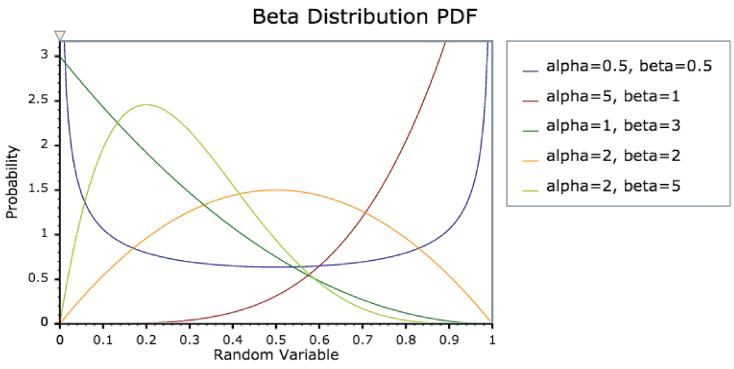


Figure 5.6: Beta distribution density for various choices of the two hyperparameters. Beta(2,2) used as a prior distribution for inclusion probability is highlighted in orange.

5.2.1.2.2 Selection results of Bayesian Group Lasso with Spike and Slab prior

Model selection criteria

For BGL-SS, there are two feasible criteria for selecting the best subset of explanatory predictors:

1. **The posterior median estimation of regression effects**: Xu et al. (2015) showed that the sub-model selected according to the posterior median estimation of regression effects has an excellent performance for both variable selection and estimation (Table 5.4).

Latent variables	Physical interpretation	Best subset of predictors
ϕ_1	deep radiocarbon	MAT
ϕ_2	topsoil radiocarbon	Land, Soil, Dif_T, MAP
ϕ_3	radiocarbon incorporation	Land, soil, Dif_T, MAP, AI, MAT

Table 5.4: The best subsets of climatic and environmental predictors for latent linear models of ϕ_1, ϕ_2 and ϕ_3 chosen according to the posterior median estimation of regression effects.

2. **The Posterior Inclusion Probability (P.I.P)**: Barbieri et al. (2004) showed that, for a linear model, the optimal predictive model is often the median probability model, which is defined as the model consisting of predictors which have overall posterior probabilities greater than or equal to 1/2 of being in a model (Table 5.5).

Latent variables	Physical interpretation	Best subset of predictors	PIP*100
ϕ_1	deep $F^{14}C$	MAT	61
		Land	53
		Soil	51
ϕ_2	topsoil $F^{14}C$	Soil	100
		MAP	78
		Dif_T	60
		Land	56
		AI	53
ϕ_3	$F^{14}C$ incorporation	Soil	97
		AI	94
		MAP	84
		Land	82
		Dif_T	72
		MAT	69
		min_P	55

Table 5.5: The best subsets of climatic and environmental predictors for latent layers ϕ_1, ϕ_2 and ϕ_3 chosen according to the Posterior inclusion Probability (PIP). The significant predictors are detected with a PIP at least equal to 0.5. The predictors highlighted in blue were the ones detected in addition to those identified by the posterior median estimation of regression effects. The predictors are ordered according to the PIP.

According to Table 5.5, the Aridity Index (AI) and the minimum precipitation (min_P) were included, in the ϕ_2 and ϕ_3 latent linear models respectively, in addition to the predictors identified for ϕ_2 and ϕ_3 in Table 5.4. The inclusion of the Aridity Index with a rather small PIP (PIP = 53) can be explained by the correlation between this predictor and the Mean Annual Precipitation (MAP) detected as significant (Figure 2.3 in Chapter 2) with both selection criteria (Pearson correlation, P.C(AI,MAT) = 0.66). The inclusion of the minimum precipitation (min_P) with a PIP equal to 55 can also be explained by the positive relationship existing between this predictor and the Aridity Index (P.C(min_P,AI) = 0.67).

The model selected according to the posterior median estimation of regression effects will be more robust to the collinearity problem than the sub-model selected according to the posterior inclusion probabilities of predictors.

Comparison of "Best BGL-SS sub-models"

Two sub-models are in competition, the one obtained using median probability criteria (Table 5.4) and the one using PIP criteria (Table 5.5). Which one to choose? In light of the Bayesian model checking and the comparison of the model criteria presented in the previous section, the most parsimonious model was chosen. The DIC was computed by testing the hierarchical non linear model (without the selection step) and considering only the predictors detected as significant with each of the BGL-SS selection criteria. The aim of this step is to readjust the estimation of regression effects by removing the predictors detected as non significant.
According to the DIC criterion, the model that best fits the dataset is the posterior median model including predictors with Posterior inclusion Probabilities (PIP) at least equal to 0.5 (the lowest recorded DIC is -2356). The Cross Validation (C.V), obtained by splitting the data into (k = 5) groups, showed no important difference between the Relative Error (R.E) computed for the sub-model selected according to the posterior median estimation of the regression effects and the Relative Error (R.E) for the model selected based on the posterior inclusion probabilities for predictors. We chose the PIP model since it has a better adjustment of data (Table 5.6).

Models	DIC	p-value of P.P.C	R.E on validation sets 5-fold C.V(%)	R.E on learning sets 5-fold C.V	Posterior coverage on validation sets(%)
Median model	-2340	0.568	36.22	11.89	95.5
PIP model	-2356	0.566	36.87	11.62	95.1

Table 5.6: Bayesian criteria comparison for the posterior median model and the model selected according to the posterior inclusion probabilities of predictors. The model with the lowest Deviance Information Criterion (DIC) is preferred to models with higher DIC. The model with a p-value of the Posterior Predictive Check (P.P.C) close to 0.5 is preferred to models with p-values close to 0 or 1. R.E refers to the Relative Error computed for both learning and validation sets.

For Bayesian Group Lasso with Spike and Slab prior, we selected the sub-model according to the Posterior Inclusion probability in Table 5.5, as it shows a better prediction power than the model selected based on the posterior median estimation of regression effects.

5.2.1.3 Bayesian Sparse Group Selection

5.2.1.3.1 BSGS model specification and choice of hyperparameters

After identifying that the soil type influences the topsoil radiocarbon dynamics by applying the Bayesian Group Lasso with Spike and Slab prior, soil scientists may wonder:

Which are the levels with significant contributions (non null effects) within the soil type that affect the topsoil radiocarbon?

The Bayesian Sparse Group Selection is constructed to handle the presence of categorical predictors in the linear model and to answer this type of question. This method performs better than the Sparse Group Lasso with Spike and Slab prior in terms of selecting the active predictors as well as identifying the active levels within the selected predictors (Chapter 4). The Bayesian Sparse Group Selection model will be applied on the latent linear models within the framework of the hierarchical non linear model proposed for soil radiocarbon dynamics. For the prior specifications, the hyperparameters are chosen as follows:

- **Prior inclusion probability:** we propose a Beta distribution, with both hyperparameters equal to 2, on prior probability for predictors and levels inclusion. This induces a distribution for the number of included variables which has a heavier tail than the binomial distribution, allowing the model to learn about the degree of sparsity.

- **The variance hyperparameter of Normal distribution for regression effects:** Zellner's g prior is proposed on the regression effect when the relative predictor is selected. The g_i^*, for i = 1, 2, 3, 4 represents the posterior mean estimation of g_i obtained from the Bayesian Full model (including all predictors). This choice was made to allow plausible values for regression effects.

BAYESIAN SPARSE GROUP SELECTION

- **Likelihood**:

  for each site $s \in \{1 : S\}$ (S = 131) and each depth $x \in \{1 : m_s\}$, the likelihood is written as:

  - $F^{14}C(s,x) \sim N_+\left(g(\phi(s),x), \sigma^2\right)$

  where, $g(\phi(s),x) = \phi_1(s) + (\phi_2(s) - \phi_1(s)) * exp\left(-\frac{x}{\phi_3(s)}\right)^{\phi_4(s)}$

- **Latent variables:**

  for each site s, the linear models are defined us:

$$\phi_i(s) \sim N_+\left(X[s,] * \beta_i, \sigma_i^2\right) \quad i = 1,2,3$$
$$log(\phi_4(s)) \sim N\left(X[s,] * \beta_4, \sigma_4^2\right)$$

- **Priors**:

  - $1/\sigma^2 \sim Gamma(0.001, 0.001)$

    for i = 1, 2, 3, 4:

  - $1/\sigma_i^2 \sim Gamma(0.001, 0.001)$

  - for the fourth latent variable ϕ_4:
    $\beta_{4,j} \sim N(0, g_4^* \sigma_4^2 (X_j' X_j)^{-1})$ for j = 1,..., P and P = 20

  - Intercepts, for i= 1, 2, 3:
    $\beta_{0,i} \sim N(0, g_i^* \sigma_i^2 (X_0' X_0)^{-1})$

  - for the two categorical predictors g = 1, 2 and for the latent layer i = 1, 2, 3:

    * **binary indicator for categorical predictor inclusion**:
      $\nu_g(i) \sim Ber(p_{predictor}(i))$
    * **binary indicator for level inclusion:**
      $\lambda_{lg}(i)|\nu_g(i) \sim (1 - \nu_g(i))\delta_0 + \nu_g(i) Ber(p_{level}(i))$
    * **predictor prior inclusion probability**:
      $p_{predictor}(i) \sim Beta(2,2)$
    * **level prior inclusion probability**:
      $p_{level}(i) \sim Beta(2,2)$

- **Priors**:
  - **regression effect of level l within the categorical predictor g:**
    $\beta_{lg}(i)|\lambda_{lg}(i)\nu_g(i) \sim (1-\lambda_{lg}(i))\delta_0 + \lambda_{lg}(i)\nu_g(i)N(0, g_i^*\sigma_i^2(X_g'X_g)^{-1})$
- for the six numerical covariates, we used the Bayesian selection method of Kuo and Mallick (1998) . This approach is based on the Stochastic Search Variable Selection introduced by George and McCulloch (1993). The spike and slab prior proposed on regression effects is replaced by a mixture model between a mass point at 0 (Dirac distribution) and a Normal distribution.
  - **regression effect for numerical predictor n = 1...,6:**
    $\beta_n(i) \sim (1-\nu_n(i)) * \delta_0 + \nu_n(i)N(0, g_i^*\sigma_i^2(X_n'X_n)^{-1})$
  - **binary indicator for numerical predictor n:**
    $\nu_n(i) \sim Ber(p_{predictor}(i))$

5.2.1.3.2 Selection results of Bayesian Sparse Group Selection

Best subset of predictors selected

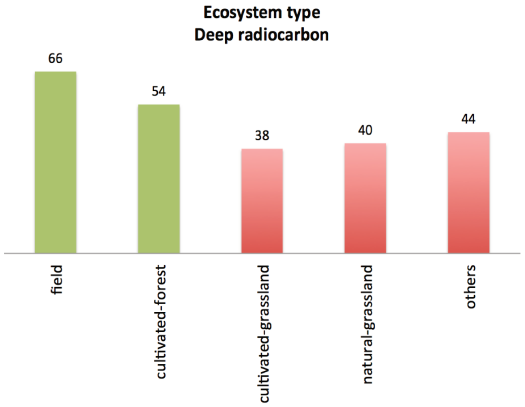
The best sub-model chosen according to the Posterior Inclusion Probability is summarized in Table 5.7:

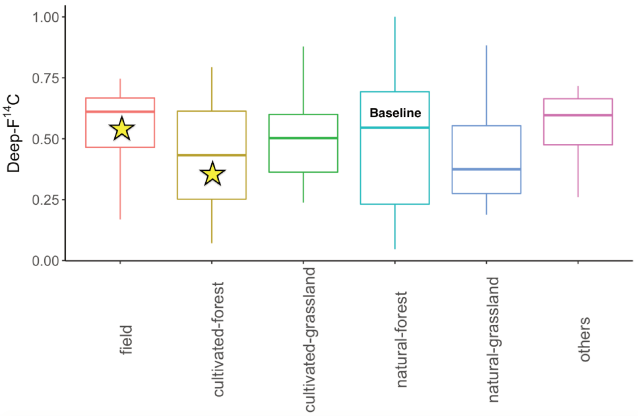
Latent variables	Physical interpretation	Best subset of predictors	PIP*100
ϕ_1	deep $F^{14}C$	Land	80
		Soil	51
		MAT	85
		MAP	54
		Dif_T	52
ϕ_2	topsoil $F^{14}C$	Land	82
		Soil	87
		Dif_T	73
ϕ_3	$F^{14}C$ incorporation	Land	73
		Soil	77
		MAP	58
		AI	77

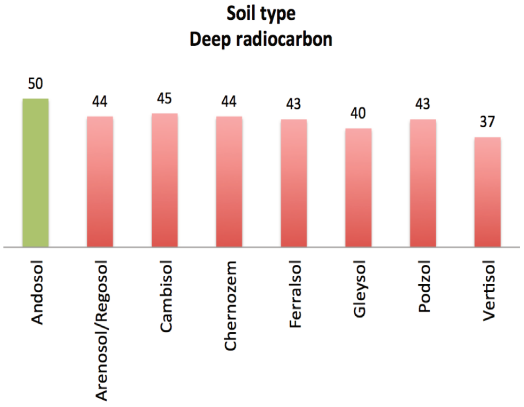
Table 5.7: The best subsets of climatic and environmental predictors for latent linear models for ϕ_1, ϕ_2 and ϕ_3 chosen according to the Posterior inclusion Probability (PIP). The significant predictors are detected with a PIP at least equal to 0.5.

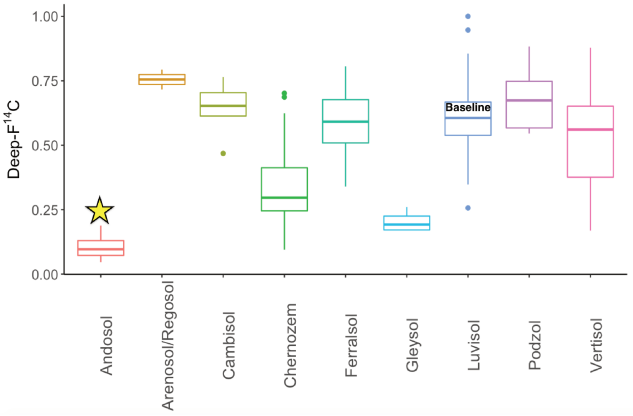
Identification of the active levels within the selected categorical predictors

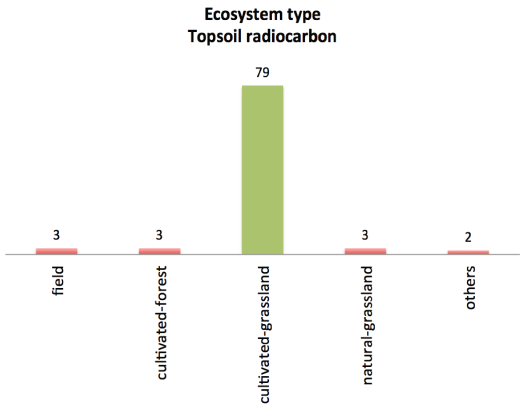
The active levels are detected with a posterior inclusion probability at least equal to 0.5. In Figure 5.7, the results of the Posterior Inclusion probability (PIP) obtained within the influential categorical predictors are presented as bars versus the real variation for deep and topsoil radiocarbon according to ecosystem and soil type shown as boxplots.

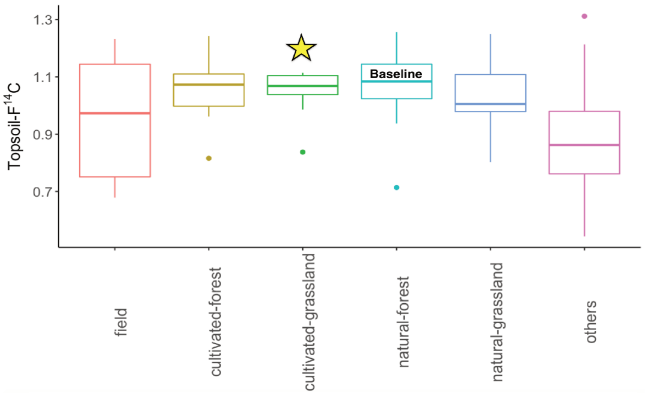

Ecosystem type
Deep radiocarbon
66
54
38
40
44
field
cultivated-forest
cultivated-grassland
natural-grassland
others


Deep-F14C
1.00
0.75
0.50
0.25
0.00
Baseline
field
cultivated-forest
cultivated-grassland
natural-forest
natural-grassland
others


Soil type
Deep radiocarbon
50
44
45
44
43
40
43
37
Andosol
Arenosol/Regosol
Cambisol
Chernozem
Ferralsol
Gleysol
Podzol
Vertisol


Deep-F14C
1.00
0.75
0.50
0.25
0.00
Baseline
Andosol
Arenosol/Regosol
Cambisol
Chernozem
Ferralsol
Gleysol
Luvisol
Podzol
Vertisol


Ecosystem type
Topsoil radiocarbon
3
3
79
3
2
field
cultivated-forest
cultivated-grassland
natural-grassland
others


Topsoil-F14C
1.3
1.1
0.9
0.7
Baseline
field
cultivated-forest
cultivated-grassland
natural-forest
natural-grassland
others

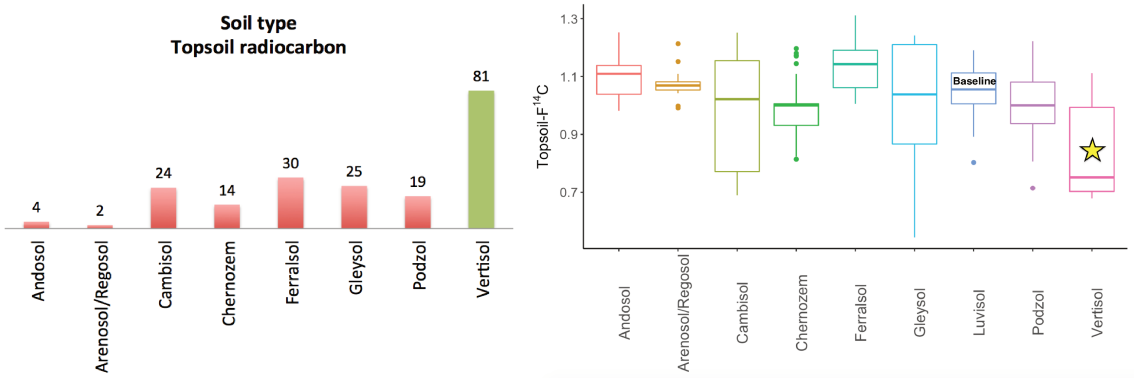


Figure 5.7: A comparison between the Posterior Inclusion Probability (given in %) for levels within the significant categorical predictors versus the real variation of radiocarbon. The green bars correspond to a Posterior Inclusion Probabilities at least equal to 0.5 (active levels) while the red ones underline the non active levels ($PIP < 0.5$). The yellow stars in the box-plots indicate the level detected as active by the Bayesian Sparse Group Selection. The boxplots for deep radiocarbon (ϕ_1) are obtained based on the profiles of the database where the depth is greater than 100 cm. In contrast, the boxplots corresponding to the topsoil radiocarbon (ϕ_2) are plotted based on the profiles of the database where the depth is smaller than 10 cm. Luvisol and the natural forest are the baseline levels for soil type and ecosystem type respectively.

For soil type, the Bayesian Sparse Group Selection detects Andosol and Vertisol with a Posterior Inclusion Portability (PIP) of 50% and 81% respectively. The boxplots of the real variation of the radiocarbon at the top and deep soil show a different behavior of Andosol and Vertisol compared to the other soil types. For ecosystem type, field and the cultivated-forest were selected with 66% and 54% for deep radiocarbone. Add to that, the only significant effect within ecosystem type was underlined by the cultivated-grassland with 79%.

5.2.1.4 Bayesian Effect Fusion using model-based clustering

5.2.1.4.1 BEF model specification and choice of hyperparameters

Bayesian Effect Fusion is a helpful method not only for selecting categorical predictors but also for merging levels within the same predictor having the same effect on the response. This method provides answers to the following question of soil scientists:

Which soil/ecosystem types have the same influence on the soil carbon behavior? Can we consider another grouping for soil/ecosystem predictors?

However, the selection performance of this approach is sensitive to the choice of some hyperparameters for the prior specifications.

Choice of hyperparameters

- **Hyperparameter vector e_0 for Dirichlet distribution**: based on the results obtained in the simulated study in Chapter 4, we fixed all components of the hyperparameter vector e_0 for the Dirichlet distribution to 0.1. This value showed a good compromise for merging levels.
- **Variance hyperparameter for the Gaussian mixture distributions:** a sensitivity analysis was conducted to choose the hyperparameter k of the variance parameter. The selection and fusion BEF performances were tested for k = 10, 50 and 100.
- **Prior inclusion probability:** As with the previous two selection approaches, a Beta(2,2) distribution is proposed on prior inclusion probability.

The BEF model for the soil radiocarbon dynamics is written as follows:

BAYESIAN EFFECT FUSION MODEL-BASED CLUSTERING

- **Likelihood**:

  for each site $s \in \{1 : S\}$ (S = 131) and each depth $x \in \{1 : m_s\}$, the likelihood is written as:

  - $F^{14}C(s,x) \sim N_+\left(g(\phi(s),x), \sigma^2\right)$

  where, $g(\phi(s),x) = \phi_1(s) + (\phi_2(s) - \phi_1(s)) * exp\left(-\frac{x}{\phi_3(s)}\right)^{\phi_4(s)}$

- **Latent variables:**

  for each site s, the linear models are defined us:

$$\phi_i(s) \sim N_+\left(X[s,] * \beta_i, \sigma_i^2\right) \quad i = 1,2,3$$
$$log(\phi_4(s)) \sim N\left(X[s,] * \beta_4, \sigma_4^2\right)$$

- **Priors**:
- $1/\sigma^2 \sim Gamma(0.001, 0.001)$

  for i = 1, 2, 3, 4:

- $1/\sigma_i^2 \sim Gamma(0.001, 0.001)$

- for the fourth latent variable ϕ_4:
  $\beta_{4,j} \sim N(0, g_4^* \sigma_4^2 (X_j' X_j)^{-1})$ for j = 1,..., P and P = 20

- Intercepts, for i= 1, 2, 3:
  $\beta_{0,i} \sim N(0, g_i^* \sigma_i^2 (X_0' X_0)^{-1})$

- for the latent variable ϕ_i i = 1, 2, 3 and the categorical covariate X_g g = 1, 2 with C_1 and C_2 levels respectively:

  - $\beta_{gl}(i) = \sum_{l=1}^{C_g} \nu_l(i) N(\mu_l(i), \psi_g(i)) + \nu_{0g}(i)\delta_0$

  - $\nu_l(i) \sim Dirichlet_{C_g+1}(e_0)$ where $e_0 \in \mathbb{R}^{C_g+1}$
    - $\nu_{0g}(i) = 0$
    - $\nu_l(i) \sim N(m_g(i), M_g(i))$
    - $\psi_g i = V_g(i)/k$ where $V_g(i) = \frac{1}{C_g - 1} \sum_{l=1}^{C_g} (\hat{\beta}_{gl}(i) - \bar{\beta}_g(i))^2$ and $\bar{\beta}_g(i) = \frac{1}{C_g} \sum_{l=1}^{C_g} \hat{\beta}_{gl}(i)$

- k and e_0 highlighted in blue need to be tuned by the user. $m_g(i)$ and $M_g(i)$ are specified according to the suggestions of Malsiner-Walli et al. (2018).
- Like the BSGS, the selection of numerical predictors is based on Kuo and Mallick (1998) approach. For i = 1, 2 , 3 and $n = 1, \ldots, 6$:
  - **regression effect for the numerical predictor n:** $\beta_n(i) \sim (1 - \nu_n(i)) * \delta_0 + \nu_n(i) N(0, g_i^* \sigma_i^2 (X_n' X_n)^{-1})$
  - **binary indicator for the numerical predictor n:** $\nu_n(i) \sim Ber(p_{predictor}(i))$

5.2.1.4.2 Selection results of Bayesian Effect Fusion model-based clustering

Results of the sensitivity analysis of the variance parameters

The Bayesian Effect Fusion was tested with three different values of k for the Gaussian mixture distribution variances: k = 10, 50 and 100. Two levels within the same categorical predictor are fused if their Posterior Fusion Probability (PFP) is at least equal to 0.5. Furthermore, a given level is fused to the baseline if the estimation of its Posterior Median Regression Effect (PMRE) estimation is exactly equal to 0. The best sub-models are identified based on both Posterior Fusion effect and Posterior Median Regression Effect.
For these three values, one sub-model is identified (Table 5.8). The lowest DIC was recorded for the BEF with k equal to 50. However, the difference in the DIC between the BEF with k = 10 (DIC = -2354) and k = 50 (DIC = -2363) is slight. A difficulty of convergence is detected for the BEF with k = 100. Even when increasing the number of iterations, there are still some parameters that do not converge according to the potential scale reduction factor defined by Gelman et al. (1992).

latent variables	k = 10			k = 50			k = 100		
	selected predictors	DIC	R	selected predictors	DIC	R	selected predictors	DIC	R
ϕ_1	MAT Soil Land	-2354	✓	MAT Soil Land	-2363	✓	MAT Soil Land	-2336	✗
ϕ_2	Land Soil			Land Soil			Land Soil		
ϕ_3	MAP AI Land Soil			MAP AI Land Soil			MAP AI Land Soil		

Table 5.8: The same sub-set of predictors is identified under the three choices of k values. The sub-model identification is based on the Posterior Fusion Probability (PFP) and the Posterior Median Regression Effect (PMRE). The Deviance Information Criterion for model fitting is given in the DIC column. The column named "R" indicates Gelman & Rubin's potential scale reduction factor for model convergence. The check-mark underlines that the convergence has been achieved while the Xmark indicates a poor model convergence.

Results of fusion of levels within the significant categorical predictors

1. **Results of fusion of levels within the ecosystem type**

The clustering of the levels of the ecosystem type identified according to the PFP and the PMRE are represented as pie charts versus the observed variation of the radiocarbon $F^{14}C$ according to the ecosystem type illustrated by boxplots (Figure 5.8).

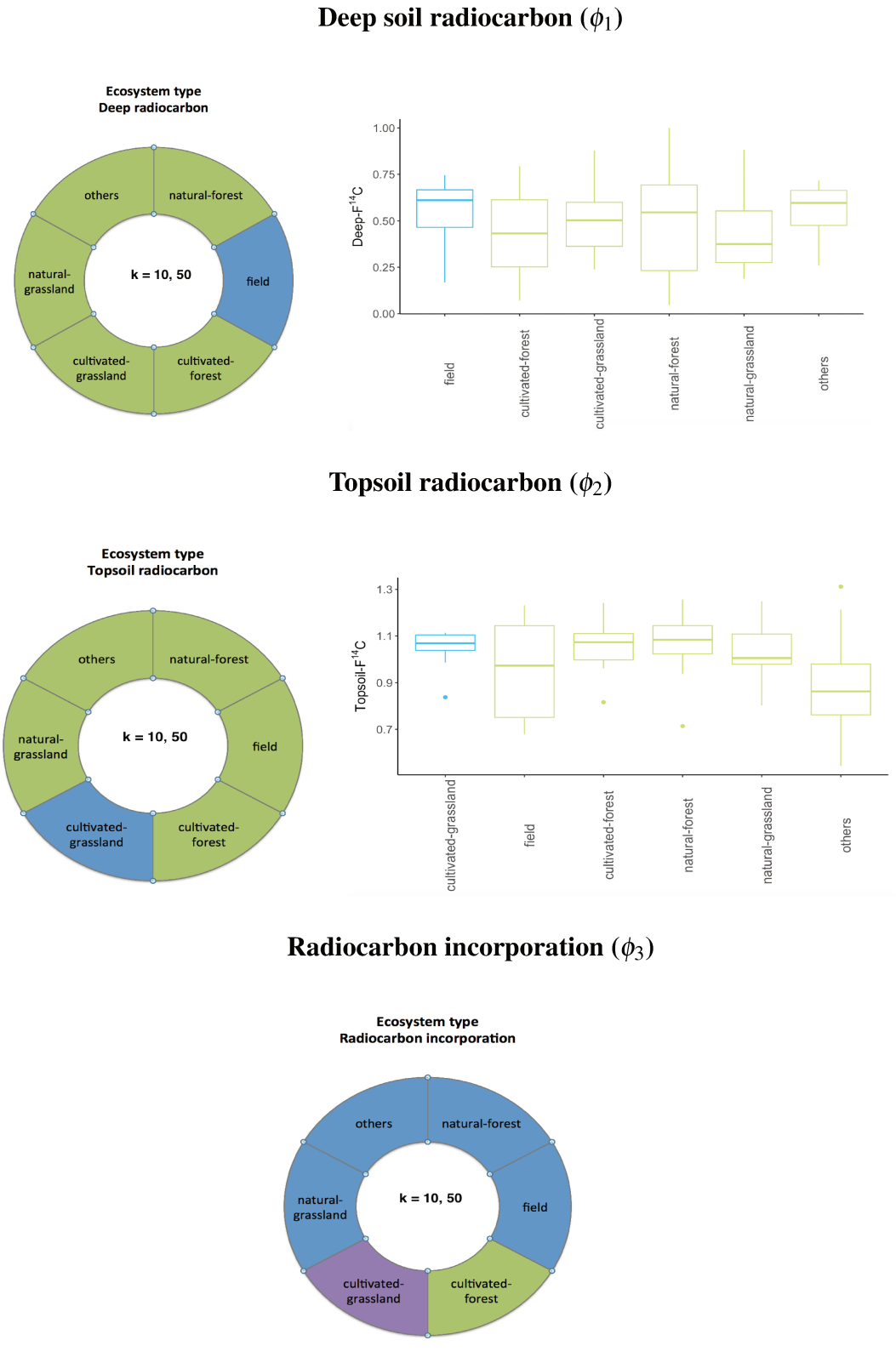
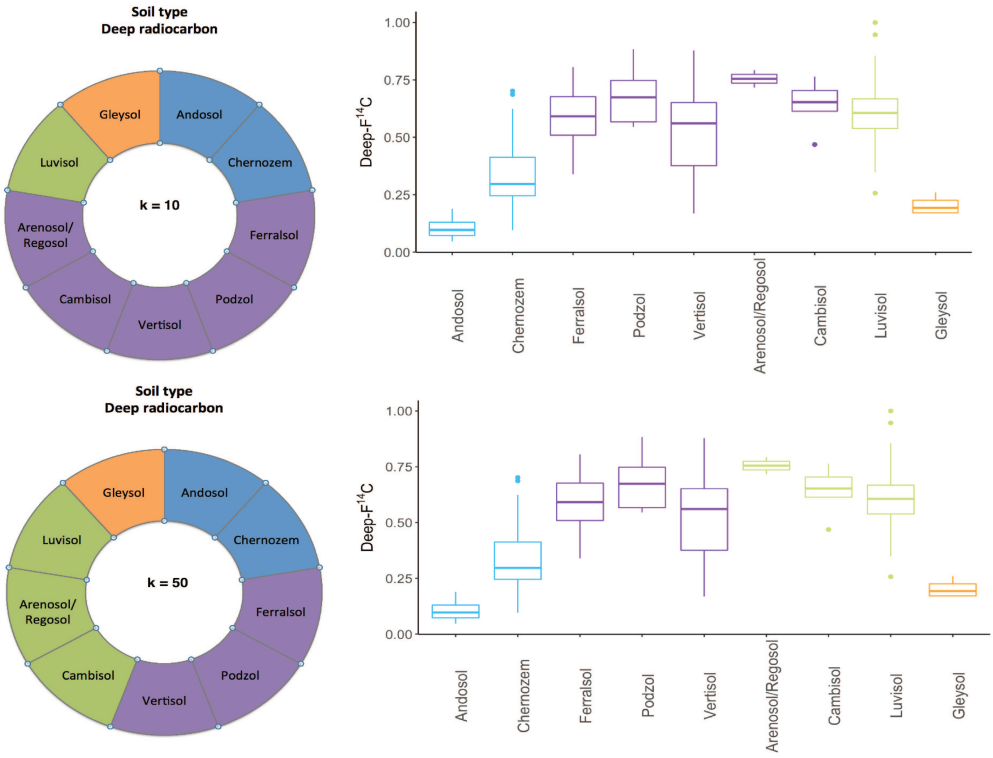


Figure 5.8: The fusion of levels for ecosystem type for the deep radiocarbon (ϕ_1), topsoil $F^{14}C$ (ϕ_2) and the $F^{14}C$ incorporation (ϕ_3) are represented by pie charts. The actual variation of the topsoil and deep radiocarbon according to the ecosystem type are illustrated by boxplots. The boxplot or piechart parts of ecosystem categories having the same color belong to the same cluster.

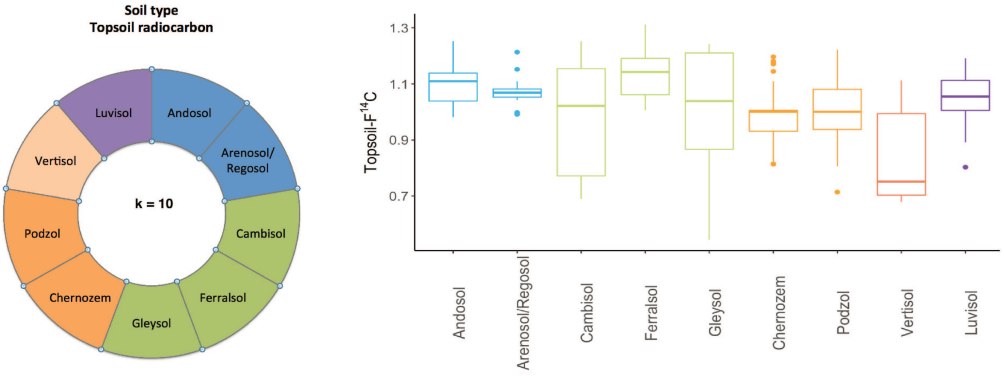
2. **Results of fusion of levels within the soil type**

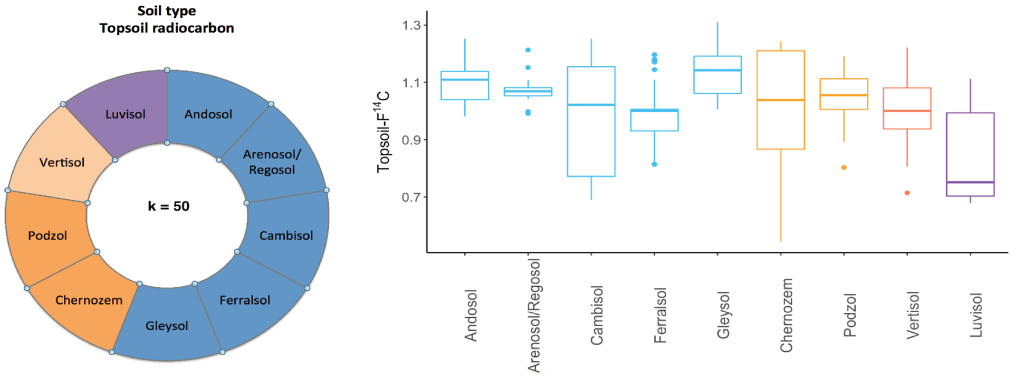
The clustering of the levels of the soil type identified according to the PFP and the PMRE are represented as pie charts versus the observed variation of the radiocarbon $F^{14}C$ according to the soil type illustrated by boxplots (Figure 5.9).

Deep radiocarbon (ϕ_1)



Topsoil radiocarbon (ϕ_2)





Radiocarbon incorporation (ϕ_3)

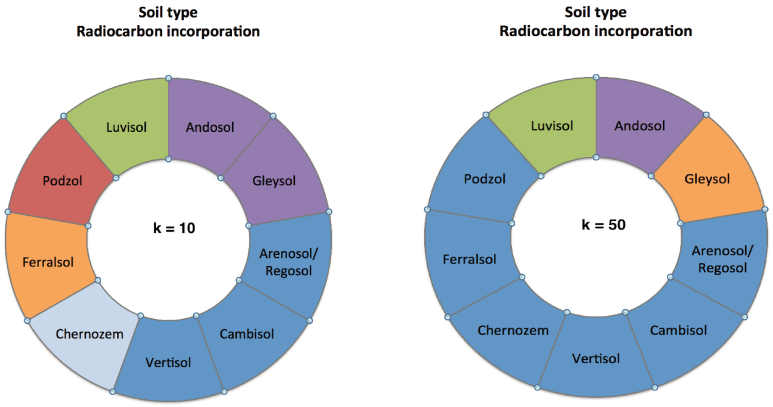


Figure 5.9: The fusion of levels for soil type for the deep radiocarbon (ϕ_1), topsoil $F^{14}C$ (ϕ_2) and the $F^{14}C$ incorporation (ϕ_3) are represented by pie charts. The real variation of the topsoil and deep radiocarbon according to the soil type are illustrated by boxplots. The boxplot or piechart parts of soil categories having the same color belong to the same cluster.

The same sub-model was identified under three different choices of Gaussian mixture hyperparameters variances k = 10, 50 and 100. The same clustering of levels of ecosystem types is obtained under the three values of k. However, the clustering of soil type levels differs from one value of k to another, the lowest DIC criterion was obtained with k = 10.

5.2.1.5 Comparison of the Bayesian selection models: Bayesian Group Lasso with Spike and Slab, Bayesian Sparse group Selection and Bayesian Effect Fusion model-based clustering

Models	DIC	p-value P.P.C	R.E on validation sets 5-fold C.V (%)	R.E on learning sets 5-fold C.V (%)	Posterior coverage on validation sets
BGL-SS (PIP model)	-2355	0.56	36.87	11.62	0.951
BSGS	-2251	0.51	36.21	11.65	0.961
BEF	-2205	0.51	35.88	11.63	0.960
Full model	-2324	0.58	37.12	15.24	0.951

Table 5.9: Comparison of the Full Bayesian model and the sub-models identified by the Bayesian selection approaches for soil radiocarbon dynamics using the Bayesian selection criteria. The model with the lowest Deviance Information Criterion (DIC) is preferred to models with a higher DIC. A p-value of the Posterior Predictive Check (P.P.C) close to 0.5 indicates a good model fitting. The model having the smallest Relative Error (R.E) on validation sets has the best predictive power. The posterior coverage of the credible intervals on the validation sets should be around 95%.

According to Table 5.9, the sub-model, identified by the BGL-SS based on the PIP , gives the best fit of the data (DIC = -2355). Among the sub-models, the difference in Relative Errors (R.E) for the three Bayesian selection methods, calculated for the validations sets, is negligible. Thus, the best sub-model is identified by the Bayesian Group lasso with Spike and Slab and according to the Posterior Inclusion Probability selection criterion.

Recap of the sub-model selected by the BGL-SS according to the Posterior Inclusion Probability

Latent variables	Best subsets of predictors
Deep $F^{14}C$ (ϕ_1)	ecosystem type (Land) soil type (Soil) Mean Annual Temperature (MAT)
Topsoil $F^{14}C$ (ϕ_2)	ecosystem type (Land) soil type (Soil) Mean Annual Precipitation (MAP) Aridity Index (AI) seasonal temperature shift (Dif_T)
$F^{14}C$ incorporation (ϕ_3)	ecosystem type (Land) soil type (Soil) Mean Annual Temperature (MAT) Mean Annual Precipitation (MAP) Aridity Index (AI) minimum precipitation (min_P) seasonal temperature shift (Dif_T)

Predictors	ϕ_1	PIP*100(ϕ_1)	P.M.E.E	ϕ_2	PIP*100 (ϕ_2)	P.M.E.E	ϕ_3	PIP*100 (ϕ_3)	P.M.E.E
Sol	✅	51	-	✅	100	-	✅	97	-
Land	✅	53	-	✅	56	-	✅	82	-
MAT	✅	61	+0.037	❌	47	+0.001	✅	69	+7.16
Dif_T	❌	44	-0.004	✅	60	-0.012	✅	72	+6.28
MAP	❌	46	+0.009	✅	78	+0.028	✅	84	+14.69
min_P	❌	42	+0.002	❌	45	-0.001	✅	55	+0.26
F^{14}Catm	❌	40	-0.004	❌	49	-0.006	❌	45	-0.05
AI	❌	42	-0.003	✅	54	-0.008	✅	94	-19.15

* P.M.E.E: Posterior Mean Effect Estimation

5.2.2 Application to soil carbon content dynamics

5.2.2.1 Full Bayesian model

5.2.2.1.1 Full Bayesian model and constraints

As for soil radiocarbon modeling, there are also constraints for soil carbon content:

1. The soil carbon content reported in g/kg is always positive. Thus, using a Normal truncated distribution or a logarithm transformation of the soil carbon content is recommended.
2. The latent variables ω_1, ω_2 and ω_3 should have positive values in accordance with their physical interpretations.

For the Bayesian inference, we consider a vague Gamma prior on the precision parameters $1/\sigma_c^2$ and $1/sd_i^2$ for $i = 1,2,3$, where both hyperparameters are equal to 0.001. With regard to regression effects, we propose a Zellner's g prior under a vague Uniform prior on g. The Full Bayesian model for soil carbon content dynamics is written as follows:

FULL BAYESIAN MODEL FOR SOIL CARBON CONTENT DYNAMICS

- **Likelihood**: for each site $s \in \{1 : S\}$ (S = 125) and each depth $x \in \{1 : m_s\}$, the likelihood is written as:
  - $C(s,x) \sim N_+\left(f(\omega(s),x), \sigma_c^2\right)$ (truncated normal distribution) or
  - $log\left(C(s,x)\right) \sim N\left(log(f(\omega(s),x)), \sigma_c^2\right)$ (logarithm transformation)

  where, $f(\phi(s),x) = \omega_1(s) + (\omega_2(s) - \omega_1(s)) * exp\left(-\frac{x}{\omega_3(s)}\right)$
- **Latent variables:**

  for each site s, the linear models are defined as:

  $$\omega_i(s) \sim N_+\left(X^*[s,] * \beta_i, sd_i^2\right) \quad i = 1,2,3$$

- **Priors**:
  - $1/\sigma^2 \sim Gamma(0.001, 0.001)$

    for i = 1, 2, 3, 4, P' = 19 and S = 125:
  - $1/sd_i^2 \sim Gamma(0.001, 0.001)$
  - $\beta_i \sim N(0, g_i sd_i^2 (X^{*'} X^*)^{-1})$ where $\mathbf{0} \in \mathbb{R}^{P'}$ and $X \in M_{S,P'}(\mathbb{R})$
  - $g_i \sim Uniform(5, 10000)$ (vague prior)

5.2.2.1.2 Results of the Full Bayesian carbon content dynamics model

In order to decide how to model the response of soil carbon contents, the additive model with truncated Normal distribution was tested against the multiplicative model with the log transformation. The Deviance Information Criterion (DIC) and the Posterior Predictive Check (P.P.C) for both models are given in Table 5.10:

Models	DIC	p-value of P.P.C
additive model (truncated Normal distribution)	7258	0.93
multiplicative model (log transformation)	2243	0.79

Table 5.10: DIC comparison between the full additive model and the multiplicative full model. The model with the lowest DIC is preferred to models with a larger DIC. The p-value of the Posterior Predictive Check is calculated according to the statistical quantity $T(C,\omega) = E(C - f(\omega, depth))$ where f is the non linear deterministic form used to model soil carbon content and C is the soil carbon content response.

The multiplicative model has the lowest Deviance information criterion (DIC = 2243). The p-value of the Posterior Predictive Check for the additive model shows that this model provides a very bad fit (p-value = 0.93). This result is in agreement with the first attempt at latent variables estimation using in optimization algorithm. In fact, the assumption of homogeneity of errors variance for an additive model was not satisfied. The residual plot is illustrated in Figure 5.10 :

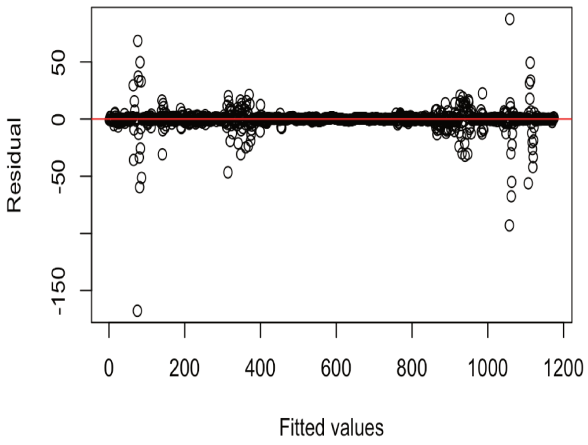


Figure 5.10: Residuals versus fitted values for soil carbon content model. Positive values for the residual mean the prediction was too low, and negative values mean the prediction was too high; 0 means the estimation was exactly correct.

In general, the posterior predictive p-value does not have a uniform distribution, under the null hypothesis that the set of parameters estimated by the model is true, but instead tends to have a distribution that clusters near 0.5. Here, the posterior predictive p-value of 0.79 highlights an overestimation of the data by the proposed statistical model. Particularly, a large uncertainty of some profiles is observed for the topsoil carbon content according to the prediction bands obtained by the statistical model. In fact, for the top 10 centimeters, the observed soil carbon content varies between 3 and 467 g/kg. This large uncertainty is also observed when linking the estimated ϕ_2, obtained by the optimization algorithm (before applying the Bayesian inference), with the climatic and environmental predictors. The estimated standard deviation is 61.2 g/kg (using the lm function in R).
Furthermore, the topsoil carbon content variation is not only large for the overall profiles of the database but also very large according to the corresponding soil type. The lack of soil expert's information about the topsoil carbon content variation and the large natural variation of topsoil carbon content according to soil type available in the SoilGrids database does not help us to better represent the uncertainties on the topsoil (see Subsection 2.1.5.6 in Chapter 2).

5.2.2.2 Bayesian Group Lasso with Spike and Slab prior

5.2.2.2.1 BGL-SS model specification and choice of hyperparameters

For the Bayesian Group Lasso with Spike and Slab (BGL-SS), we followed the same concept of prior specification of the BGL-SS used for the soil radiocarbon dynamics. Thus, a Beta prior is proposed for the prior inclusion probability with both hyperparameters equal to 2.

5.2.2.2.2 BGL-SS selection results

Best sub-models selection criteria

Two Bayesian selection criteria can be used to choose the best sub-predictors for soil carbon content dynamics. We recall that the first criterion is based on choosing the sub-model according to the posterior median regression effect estimation while the second criterion consists in including in the model all the predictors with a Posterior Inclusion probability (PIP) at least equal to 0.5. The BGL-SS selection results, in latent linear models, for both Bayesian selection criteria are summarized in Table 5.11:

Latent variables	Physical interpretation	Median model	PIP model	
		best sub-predictors	best sub-predictors	PIP
ω_1	Deep C	Land	Land	71
		Soil	Soil	71
		MAP	MAP	63
			AI	50
ω_2	Topsoil C	Soil	Soil	100
		MAT	MAT	100
		Dif_T	Dif_T	70
ω_3	C incorporation	Soil	Soil	100
		Dif_T	Dif_T	62
		MAP	map	56

Table 5.11: The best sub-predictors detected according to the two Bayesian selection criteria used for the Bayesian Group Lasso with Spike and Slab prior. The sub-predictors selected according to the posterior median estimation of regression effects are indicated in the column called "Median model". The sub predictors detected according to the Posterior inclusion Probability (PIP) are represented in the column called "PIP model". This column also contains the PIP for the significant predictors. The symbol C in the second column refers to the carbon content.

Readjustment and best sub-models comparison

In order to better estimate the regression effects, the best sub-models based on the posterior median estimation of regression effects and the posterior inclusion probability were readjusted by removing the redundant predictors. The Deviance Information Criterion, for both sub-models, is given in Table 5.12.

Best sub-models	DIC
Median model	2072
PIP model	2078

Table 5.12: The Deviance Information Criterion (DIC) for the two sub-models detected by the Bayesian Group Lasso with Spike and Slab. The Median model is based on the posterior median estimation of regression effects while the PIP model is based on the Posterior Inclusion Probability (PIP) for predictors.

According to Table 5.12, the best model chosen with respect to the lowest DIC is the Median model (DIC = 2072). Furthermore, the two sub-models can be considered as a step of Stepwise regression. In fact, the only

difference between the two sub-models is that the PIP model considers, in addition, the Aridity Index (AI) as a significant predictor for the ω_1 latent linear model with a Posterior Inclusion Probability equal to the selection threshold 0.5. An increase in DIC is observed for the PIP model after adding the Aridity Index (AI) (DIC = 2078). Thus, the final sub-model kept is the Median model.

5.2.2.3 Bayesian Sparse Group Selection

5.2.2.3.1 BSGS model specification and choice of hyperparameters

The Bayesian Sparse Group Selection (BSGS) model has the same structure as the BSGS proposed for the soil radiocarbon modeling with the exception that for the soil carbon content three latent variables are defined instead of four. This approach provides a probabilistic judgment about the inclusion of categorical predictors as well as levels. The Bayesian method of Kuo and Mallick (1998) was used to select the numerical predictors. For the prior specification, a Beta distribution with both hyperparameters equal to 2 was considered for the prior inclusion probability for categorical predictors as well as for levels. If the categorical predictor is significant, the regression effect is generated from a Zellner's g prior where the g value for each latent linear model is replaced by the posterior mean estimation of g obtained from the Full Bayesian model.

5.2.2.3.2 BSGS selection results

The best sub-model identified by the BSGS

In the best sub-model, all the predictors have a Posterior Inclusion Probabilities (PIP) at least equal to 0.5. The best sub-predictors for each latent linear model are summarized in Table 5.13.

Latent variables	Physical interpretation	Best sub-predictors	PIP $>$ 0.5 (%)
ω_1	Deep carbon content	Land	82
		Soil	83
		AI	51
		Dif_T	60
ω_2	Topsoil carbon content	Land	100
		Soil	100
		MAT	100
		MAP	64
		min_P	58
		AI	56
ω_3	Carbon content incorporation	Land	100
		Soil	100
		MAT	60
		MAP	76
		min_P	62
		AI	62
		Dif_T	82

Table 5.13: The best sub-predictors for each latent linear model identified according to the Posterior Inclusion Probability (PIP). A predictor is considered as significant if its PIP is at least equal to 0.5.

The active levels identified within the significant categorical predictors

The Posterior Inclusion Probability of levels within the significant categorical predictors are illustrated by histograms in Figure 5.11.

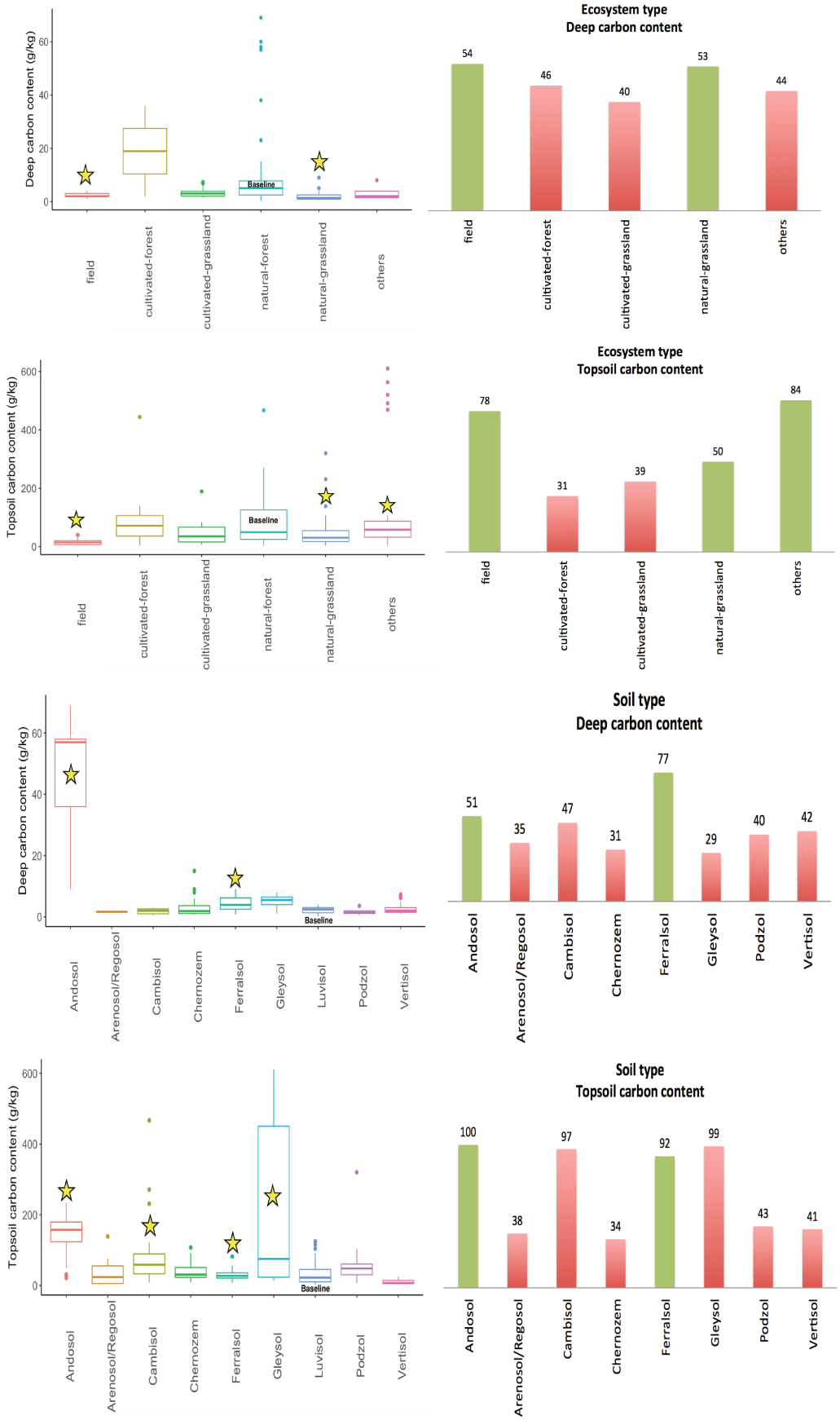


Figure 5.11: The barcharts (right panels) illustrate the Posterior Inclusion Probability (PIP) for soil and ecosystem type in the ω_1 and ω_2 latent linear models. The box-plots (left panels) illustrate the real variation of deep and topsoil soil carbon content according to soil and ecosystem types respectively. The yellow stars indicate the levels detected as active by the Bayesian Sparse Group Selection. Luvisol and natural-forest are the baseline for the soil type and ecosystem type respectively.

5.2.2.4 Bayesian Effect Fusion model-based clustering

5.2.2.4.1 BEF model specification and choice of hyperparameters

Bayesian Effect Fusion is sensitive to the choice of hyperparameters as seen in Chapter 4. Here, we fixed all the components of the vector e_0 for the Dirichlet distribution to 0.1. The BEF was tested under three different values of $k = 10, 100$ and 120.

5.2.2.4.2 BEF selection results

The results of the sensitivity analysis are summarized in Table 5.14.

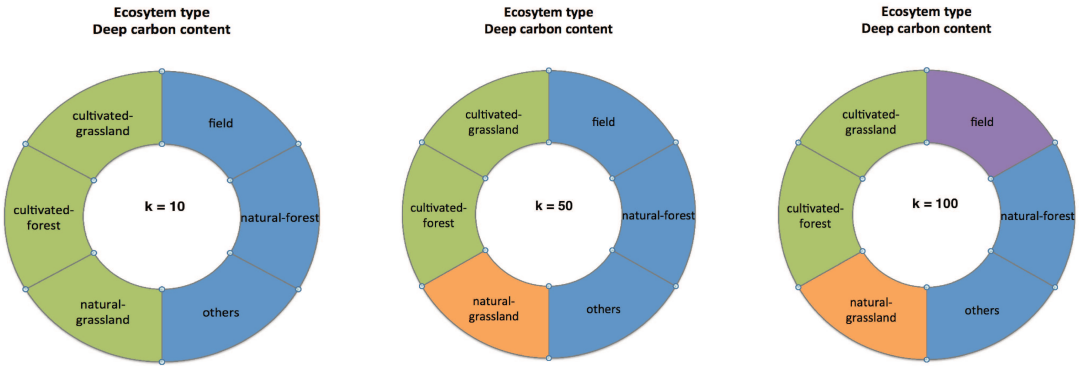
latent variables	k = 10			k = 50			k = 100		
	selected predictors	DIC	R	selected predictors	DIC	R	selected predictors	DIC	R
ω_1	Soil Land	2051	✓	Soil Land	2045	✓	Soil Land	2053	✓
ω_2	Land Soil MAT MAP			Land Soil MAT			Land Soil MAT MAP		
ω_3	Land Soil MAP Dif_T			Land Soil MAP Dif_T AI			Land Soil MAP Dif_T		

Table 5.14: The best sub-set of predictors is identified for each choice of k. The Deviance Information Criterion for model fitting is given in the column DIC. The column named "R" indicates Gelman & Rubin's convergence. The check-mark underlines that convergence has been achieved while the Xmark indicates a difficulty in achieving model convergence. The blue Xmark indicates a poor convergence.

Fusion models	DIC	p-value (P.P.C)
Model with k = 10, 100	2057	0.81
Model with k = 50	2076	0.81

Results of fusion levels with the significant categorical predictors

Results for levels of ecosystem type



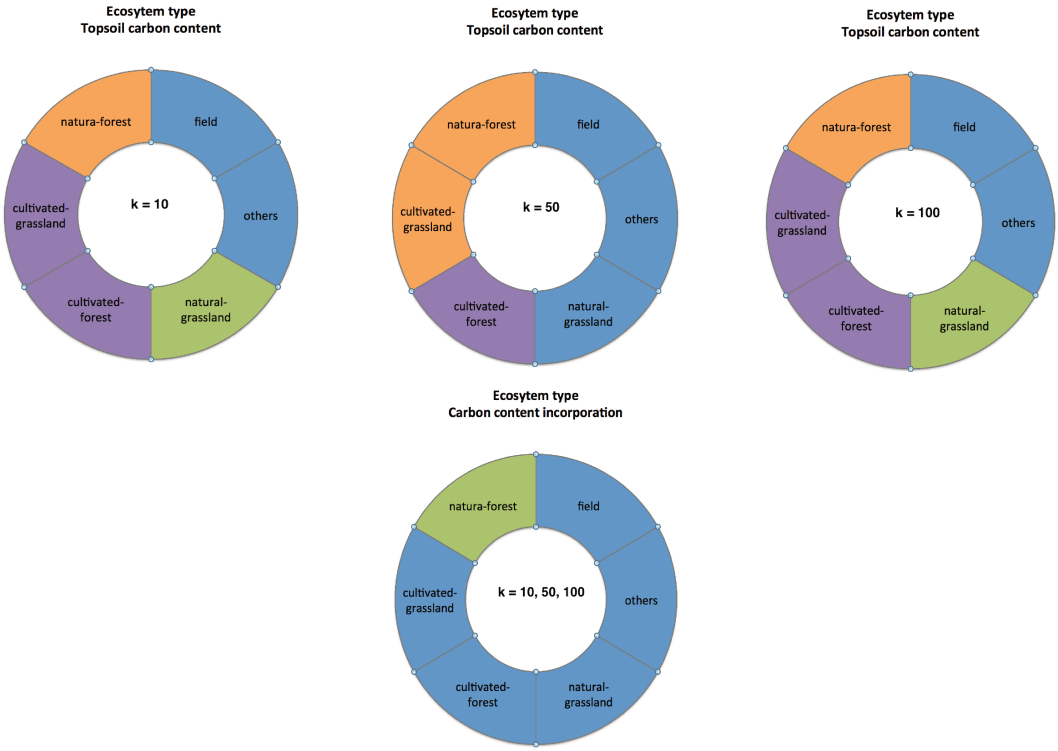
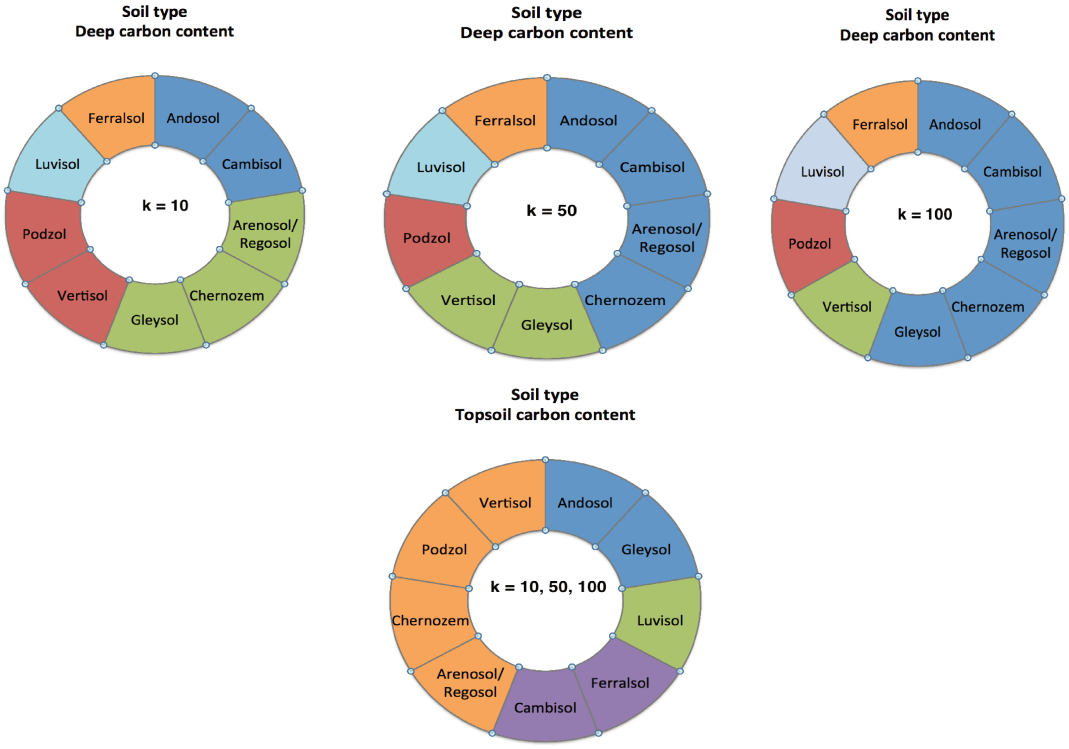


Figure 5.12: The results of the ecosystem type levels fusion for each of the three latent variables ω_1 (deep carbon content), ω_2 (topsoil carbon content) and ω_3 (carbon content incorporation). The fusion of levels is based on the Posterior Fusion probability (PFP) for pairs of levels and the posterior median for regression effects. Two levels are fused together if their PFP is at least equal to 50%.

2- Results for levels of soil type



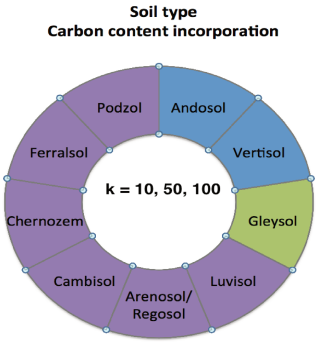


Figure 5.13: The results of the soil type levels fusion for each of the three latent variables ω_1 (deep carbon content), ω_2 (topsoil carbon content) and ω_3 (carbon content incorporation). The fusion of levels is based on the Posterior Fusion probability (PFP) for pairs of levels and the posterior median for regression effects. Two levels are fused together if their PFP is at least equal to 50%.

For the topsoil radiocarbon, the same clusters are identified under the different values of k = 10, 50 and 100 (Figure 5.13).

5.2.2.5 Comparison of the Bayesian selection methods for soil carbon content dynamics

Models	DIC	P.P.C	Median R.E on learning sets 5-fold C.V (%)	Median R.E on validation sets 5-fold C.V (%)	Posterior coverage on validation sets
BGL-SS (posterior median)	2072	0.77	61	167	0.97
BSGS	2083	0.80	76	171	0.98
BEF	2057	0.81	76	170	0.98
Full model	2091	0.79	67	145	0.95

Table 5.15: Comparison of the Full Bayesian model and the sub-models identified by the Bayesian selection approaches for soil carbon content dynamics using the Bayesian selection criteria. The model with the lowest Deviance Information Criterion (DIC) is preferred to models with a higher DIC. A p-value of the Posterior Predictive Check (P.P.C) close to 0.5 indicates a good model fitting. The model having the smallest Relative Error (R.E) on validation sets has the best predictive power. The posterior coverage of the credible intervals on the validation sets should be around 95%.

For each profile and at each depth, the sum of the Relative Error (R.E), which compares the absolute error relative to the observed value, was computed. In order to obtain an estimation for the overall profiles, we computed the median of the Relative Errors for all studied sites. Here, the Relative Error exceeds 100%. Theoretically, the relative error can be any size at all, including more than 100%. For example, if we expect a soil carbon content to be 5 g/kg but it was measured as 15 g/kg, this gives a Relative Error of 200%. This result was expected since the posterior predictive p-value was higher than 0.5.

Better adjustment is noted for the Bayesian Effect Fusion model-based clustering (BEF) (the lowest DIC recorded: 2057). However, the best prediction indicated by the lowest Median Relative Error on the validations sets is obtained with the Full Bayesian model.

Recap of the sub-model selected by Bayesian Effect Fusion

Latent variables	Best subsets of predictors
Deep carbon content (ω_1)	ecosystem type (Land) soil type (Soil)
Topsoil carbon content (ω_2)	ecosystem type (Land) soil type (Soil) Mean Annual Precipitation (MAP) Mean Annual Temperature (MAT)
carbon content incorporation (ω_3)	ecosystem type (Land) soil type (Sol) Mean Annual Precipitation (MAP) seasonal shift (Dif_T)

Predictors	ω_1	PIP*100 (ω_1)	P.M.E.E	ω_2	PIP*100 (ω_2)	P.M.E.E	ω_3	PIP*100 (ω_3)	P.M.E.E
Sol	✅	-	-	✅	-	-	✅	-	-
Land	✅	-	-	✅	-	-	✅	-	-
MAT	❌	24	-0.05	✅	100	-64.96	❌	47	+1.51
Dif_T	❌	24	+0.06	❌	31	-0.85	✅	67	-6.01
MAP	❌	21	+0.03	❌	50	+8.50	✅	62	+5.55
min_P	❌	26	+0.12	❌	38	+3.68	❌	41	-1.11
AI	❌	27	-0.03	❌	38	-4.18	❌	49	-2.73

* P.M.E.E: Posterior Mean Effect Estimation

5.3 Physical interpretations of climatic and environmental predictors selected as significant

5.3.1 Atmospheric $F^{14}C$ of the sampling year ($F^{14}C$atm)

Reminder

Atmospheric $F^{14}C$ of the sampling year was not selected as an influential predictor for any of the latent variables. Atmospheric radiocarbon only makes sense for the modeling of the $F^{14}C$ profile. It is not considered in the soil carbon content dynamics statistical study.

The atmospheric ^{14}C concentration has not always been constant with time. In addition to the natural variation of the atmospheric $F^{14}C$, two anthropogenic activities disrupted the ^{14}C cycle. The first change in the ratio of the atmospheric concentrations (decrease in the atmospheric $F^{14}C$) was caused by the admixture of large amounts of fossil-fuel which does not contain ^{14}C. Added to that, in the late 1950s and early 1960s, the atmospheric radiocarbon increased due to the massive introduction of ^{14}C from atmospheric nuclear testing (Hua et al., 2013). Other earth carbon reservoirs such as vegetation and soil, have undergone similar ^{14}C anthropogenic changes. For example, a profile sampled in 1965 is richer in topsoil radiocarbon than a profile sampled in 1990.

- **Impact of the sampling year of the atmospheric $F^{14}C$ ($F^{14}C$atm)**

  The Posterior Inclusion Probability of the atmospheric $F^{14}C$ predictor, for topsoil $F^{14}C$ (ϕ_2), is estimated at 49%. Even very close to the selection threshold (PIP = 50%), the atmospheric $F^{14}C$ is not detected as a significant predictor to explain the topsoil $F^{14}C$. This was highly unexpected.
  However, on returning to the original data from the database, we do indeed not see any clear relationship between topsoil $F^{14}C$ and atmospheric $F^{14}C$ (Figure 5.14). This non-selection of $F^{14}C$atm can be explained by the bias that derived from the database concerning the sampling year of the profiles. In fact, 53% of the profiles were sampled in the nineties (see Figure 2.7 in Chapter 2). The database is unbalanced with respect to the number of profiles sampled before and after the influence of bomb ^{14}C.

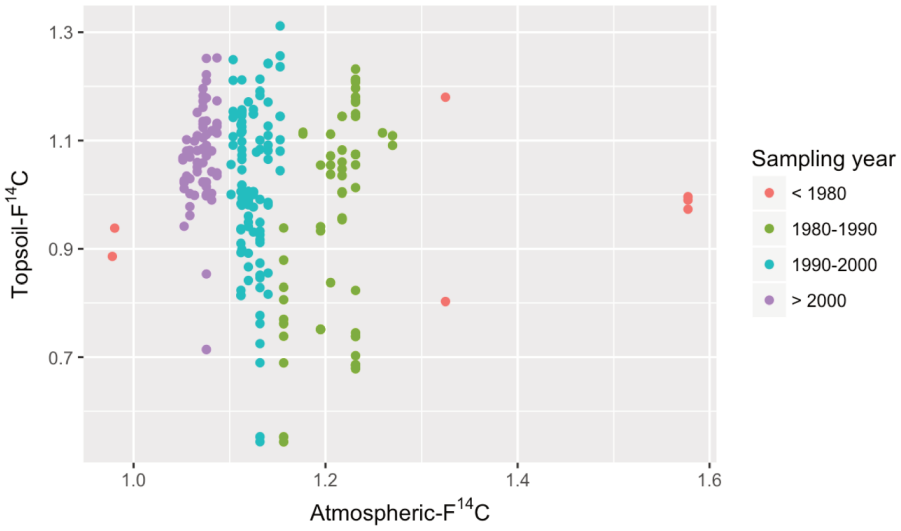


Figure 5.14: Distribution of the topsoil (less than 10 cm depth) $F^{14}C$ from the database versus the atmospheric $F^{14}C$. Colors highlight sampling years grouped into four periods.: before 1980, [1980-1990], [1990-2000] and after 2000.

The non detection of atmospheric radiocarbon as an influential predictor, for the topsoil radiocarbon, is due to a poor representation of the potential sampling years in the database, resulting in an overrepresentation of profiles collected in the 1990s.

5.3.2 Mean Annual Temperature (MAT)

Reminder

Mean Annual Temperature (MAT) was selected as an influential predictor for deep $F^{14}C$ (ϕ_1) and $F^{14}C$ incorporation depth (ϕ_3) of the $F^{14}C$ profile and for the topsoil (ω_2) of the carbon content profile

It is globally assumed that a cooler temperature is associated with slower decomposition and increases the mean residence time of soil carbon. The smallest change in the soil carbon content may have a large impact on the concentration of CO_2 in the atmosphere (Trumbore et al., 1996). The modalities of temperature impact on the decomposition rate of soil organic matter remain an interesting topic of discussion. A study done by Giardina and Ryan (2000) suggested that the recalcitrant carbon is not sensitive to temperature variation. In contrast, Fierer et al. (2005) suggested that the non-labile organic matter is more sensitive to temperature than the labile pool. Fang et al. (2005), on incubated soils under changing temperature, found similar results. Likewise, Lefevre et al. (2014) highlighted, on long-term (to 79 years) bare fallow experiments, a strong relationship between the residence time of carbon organic matter and the temperature sensitivity of its mineralization: the more stable the organic matter, the more sensitive to temperature the organic matter will be. And finally, Conen et al. (2006) pointed out that recalcitrant and labile pools have a similar temperature sensitivity.

A recent study done by Yan et al. (2017) also took a position in this debate, pointing out different fates between top and deep soil. This study was based on a sequential temperature (8°C to 28°C) changing method applied on cultivated fields in China. Results showed that the average SOC decomposition rate was 59% to 282% higher in the topsoil than in the subsoil layer. In contrast, the temperature sensitivity values in the topsoil layer were significantly lower than those in the subsoil layer.

The increasing temperature in humid climates increases both plant growth and decomposition of soil organic matter. However, the relative increase in the decomposition rate of organic matter remains greater than the net primary production (Oades, 1988). In a laboratory study, Hagerty et al. (2014) showed that the microbial turnover accelerates with temperature while the growth efficiency is not sensitive to temperature changes.

It is certain that in order to extract the sole effect of temperature on soil carbon content and dynamics, it would be necessary to work with several temperatures on the same environmental conditions (soil type, ecosystem, vegetation cover, etc.) which was not always applicable for the studies carried out. Several parameters influencing soil carbon are thus mixed and it is difficult to deconvolve the signal. Furthermore, the sensitivity of soil carbon content to temperature is often studied by confronting soil carbon results with Mean Annual temperature (MAT) or the seasonal shift of temperature not immediate temperature (Smith et al., 2008).

In this debate, the statistical results obtained on our meta-analysis of radiocarbon and soil carbon content can provide elements for decision making

5.3.2.1 Impact of the Mean Annual Temperature (MAT) on the mean residence time ($F^{14}C$)

- **Impact of the Mean Annual Temperature (MAT) on the topsoil and deep $F^{14}C$**

The Mean Annual Temperature (MAT) is detected as significant with a posterior inclusion probability of 61% for the deep radiocarbon activity, whereas it is not detected as an influential predictor for the topsoil radiocarbon (Table 5.5). Nevertheless, the MAT for the topsoil radiocarbon was associated to a PIP of 47%, which is not very far from the selection threshold of 50%.
For **deep soil, the posterior mean estimation of the effect of the MAT is estimated at 0.03** (with 91% of chance of being positive). In other words, an increment of 1°C in the MAT will result on average in an increase of the deep $F^{14}C$ of 0.03 (standard deviation of 0.03). For example, let us assume that $F^{14}C = 0.3$ (^{14}C age of about 9670 yr BP) is the deep soil radiocarbon measurement under a given soil type and land vegetation cover. An increment of 1° C would increase, on average, the $F^{14}C$ by 0.33, *i.e.* will result in a decrease of the mean ^{14}C age from 9670 yr BP to 8910 yr BP (more details about the link between $F^{14}C$ and ^{14}C age are given in Appendix 7.1.

This unexpected result, *i.e.* MAT as a significant predictor for deep soil but not for topsoil, is confirmed by the database analysis. The confrontation of topsoil and deep soil $F^{14}C$ with Mean Annual Temperature (Figure 5.15) highlights a positive trend between $F^{14}C$ and MAT for deep soil whereas no clear signal can be discerned for topsoil. A younger soil radiocarbon is indeed recorded for higher MAT (Figure 5.15b) for the deep soil $F^{14}C$, *i.e.* for depth at least equal to 100 cm. In contrast, for the topsoil (depth not exceeding 10 cm) only a minor change in soil $F^{14}C$ is observed with an increase in MAT (Figure 5.15a).

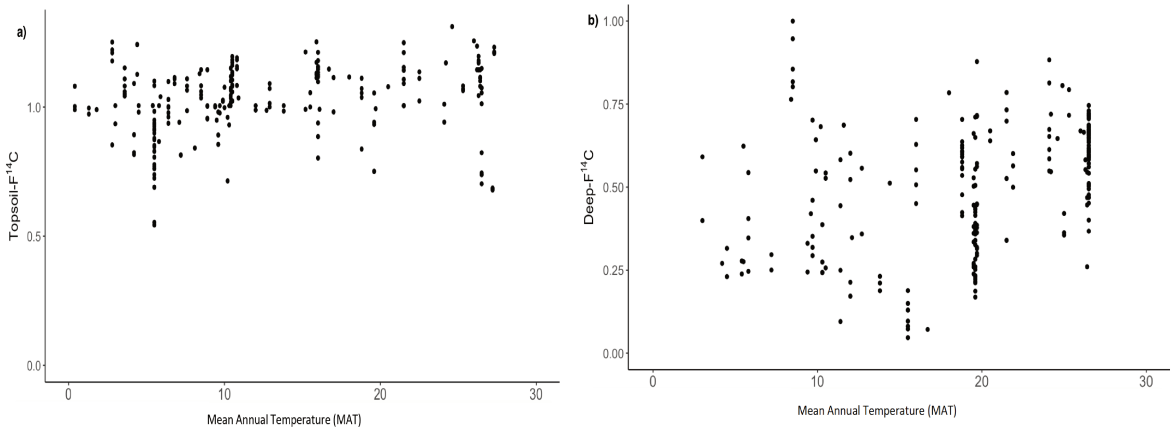


Figure 5.15: Topsoil (depth $\leq$ 10 cm, panel a) and deep (depth $>$ 100 cm, panel b) $F^{14}C$ from the database versus the Mean Annual Temperature (MAT) for the database radiocarbon profiles.

In panel b of Figure 5.15, the set of points, framed by a green circle, shows a different fate to that of the global trend. All the particular values of $F^{14}C$ measurements correspond to a Luvisol profile from the same site sampled in Germany under the Wohldorf forest (Becker-Heidmann and Scharpenseel, 1986). This profile shows a surprising level between 100 and 112 cm depth with high $F^{14}C$ values that do not fit with the general trend of the profile. The authors attributed this feature to the percolation of young organic carbon from the top that stops its course at the boundary between B and C horizons. The same feature is found at the A/B boundary as well but as it is higher in the profile, it does not have the same impact. As Becker-Heidmann and Scharpenseel (1986) sampled soil profiles at a very high resolution (typically a 2 cm step), this very particular

feature is overrepresented in our database but as the selection approach deals with profile shape and not with the individual points, the impact of their presence in the final result is not that high. Nevertheless, this profile might have biased the evaluation of the impact of the MAT predictor, which might have been a bit higher than the current 0.03 ± 0.03 $F^{14}C$ unit/°C without the German forest deep points.

- **Impact of the Mean Annual Temperature (MAT) on $F^{14}C$ incorporation depth**

  MAT was detected as significant for $F^{14}C$ incorporation (ϕ_3) with a posterior inclusion probability of 69%. The posterior mean of the effect of the MAT, expressed in cm, is **estimated at -6 cm** with 10 cm of standard deviation (74% chance of having a negative effect). The signal of -6±10 cm/°C is not that clear but it might be the expression of the loss of some decades old organic carbon (high $F^{14}C$ as marked by the bomb peak) at constant depth (Figure 5.16, 20 cm deep, $F^{14}C$ shifts from 0.59 to 0.43).

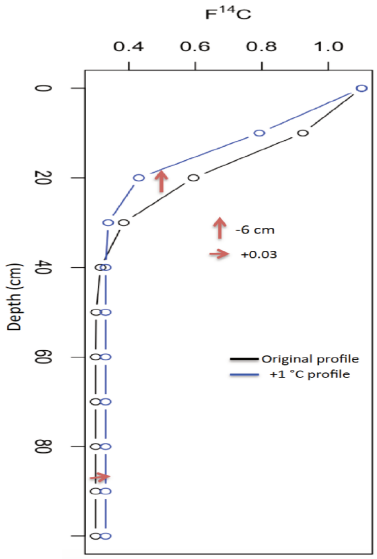


Figure 5.16: The impact of the increment of 1°C in the Mean Annual Temperature (MAT) on the deep $F^{14}C$ (ϕ_1) and the $F^{14}C$ incorporation (ϕ_3). The profile is plotted for $\phi_1 = 0.3, \phi_2 = 1.1, \phi_3 = 20$ and $\phi_4 = 2$. The modified profile is given in the blue curve.

5.3.2.2 Impact of the Mean Annual Temperature (MAT) on the topsoil organic carbon content

According to the best sub-model identified by the Bayesian Effect Fusion on the soil carbon content profile, the **MAT predictor was considered as significant only for the topsoil carbon content**. This selection is associated to a PIP of 100%. The posterior mean **effect of the MAT is estimated at -63 g/kg** with 15 g/kg of standard deviation (100% chance of having positive values). This means that an increment of 1°C in the MAT decreases, on average, the topsoil carbon content by 63 g/kg. This observation can be linked to a higher vegetation primary production under higher temperature, considering the MAT range of the database that excludes extreme climates such as desert. The higher primary production should increase the carbon input to the soil.

In conclusion, an increase in MAT by 1 °C might lead to i– a significant decrease in the topsoil carbon content without affecting the $F^{14}C$ and thus the mean residence time, ii- a highly likely increase in deep soil $F^{14}C$ reflecting a lower mean residence time and thus more rapid organic matter turnover under higher temperature, iii- a likely loss of some decades old organic carbon at mid-depth. So an increase in MAT has an impact on both the short-term as it results in a decrease of topsoil carbon content and the long-term as highlighted by the change in the mean residence time (loss of centennial to millennial carbon in depth and loss of decadal carbon at mid-depth).

5.3.3 Mean Annual Precipitation (MAP)

Reminder

Mean Annual Precipitation (MAP) was selected as an influential predictor for topsoil (ϕ_2) and incorporation depth (ϕ_3) of the $F^{14}C$ profile and for the topsoil (ω_2) and incorporation depth (ω_3) of the carbon content profile.

Precipitation was identified as a possible key driver for soil radiocarbon and carbon dynamics, because it can translocate fresh, young organic carbon from the surface to the subsoil. Deng et al. (2016), based on a three year field experiment in subtropical forests in China, showed that an increase in precipitation results in a slight increase in soil respiration. This result may, however, be specific to forests since they are characterized by quite low soil moisture (Zhou et al., 2006). However, Han et al. (2018) showed that soil organic carbon stocks significantly increase along precipitation gradient in all vegetation types except woodland.
Indeed, an increase in precipitation is generally associated with high vegetation growth rates and therefore high inputs of organic carbon into soil (Liu et al., 2011; Wang et al., 2010).

5.3.3.1 Impact of the Mean Annual Precipitation (MAP) on the mean residence time ($F^{14}C$)

- **Impact of the Mean Annual Precipitation (MAP) on the topsoil $F^{14}C$**

Mean Annual Precipitation (MAP) was detected as an influential predictor for the topsoil $F^{14}C$ with a posterior inclusion probability of 78%. The posterior mean estimation of the effect of MAP is **estimated at $\pm$ 0.002** (standard deviation of 0.02) with 53% chance of having a positive effect **(Figure 5.17)**. +0.002 $\pm$ 0.02 $F^{14}C$ unit/mm means that MAP effectively impacts topsoil $F^{14}C$ but it might sometimes be towards more negative values and sometimes towards more positive values. The impact of MAP is either not clear in reality, or not clearly expressed in the database or not captured by the mathematical approach.

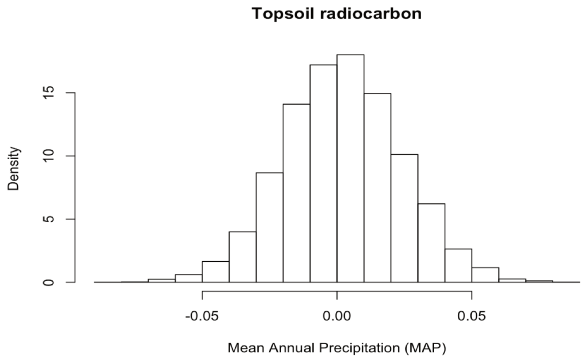


Figure 5.17: Histogram of the posterior distribution of the Mean Annual Precipitation (MAP) effect on the topsoil radiocarbon (ϕ_2).

More in-depth analysis of the database underlines that recorded topsoil $F^{14}C$, sampled from a depth not exceeding 10 cm, shows a general positive trend with MAP (Figure 5.18). This trend may not have been detected as clearly as it should have been by our mathematical approach due to the series of values belonging to 7 profiles at 1600 mm of MAP (red circle in Figure 5.18). These values of the 7 profiles correspond to the same site located at the Moor House Nature Reserve (MHNR) in the United Kingdom (UK) (Bol et al., 1996; Huang et al., 1996, 1999). Moor house Nature Reserve Habitats include exposed summits, extensive blanket peatlands, upland grasslands, pastures, hay meadows and deciduous woodland. Analysis of the soil

map (https://catalogue.ceh.ac.uk/maps≠layers/b36357bd-988c-41fa-a3a8-3b21cef5f0b6) shows that most soil types are blanket bog associated to some peaty Gley and peaty Podzol. The "peaty" qualifier was not detected at the first reading of the database but here clearly shows that it is the main feature of the soil. As we wished to focus on mineral soil, excluding peat and permafrost, we should have removed the profiles from this reserve for our study.

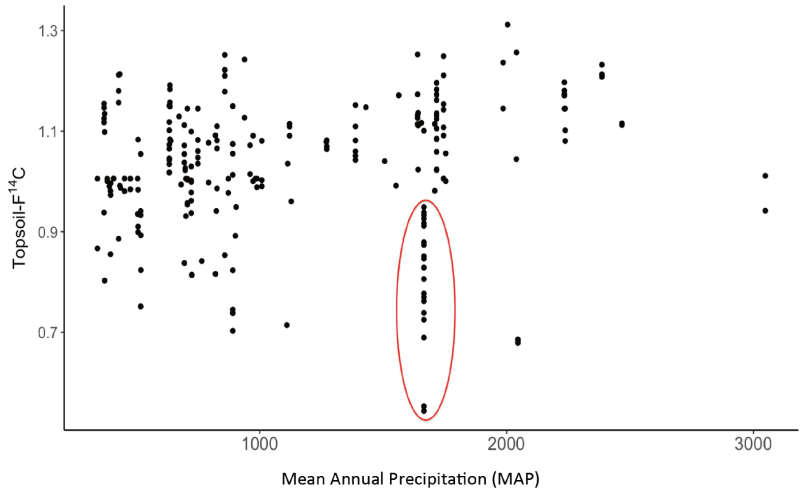


Figure 5.18: Distribution of the topsoil radiocarbon (less than 10 cm depth) $F^{14}C$ from the database versus the Mean Annual Precipitation (MAP).

To illustrate the impact of MAP on topsoil $F^{14}C$, let us assume a topsoil $F^{14}C$ equal to 1.1. An increment of 1 mm in the MAP leads to an increase, on average, of 0.002 $F^{14}C$ unit, shifting the topsoil $F^{14}C$ from 1.1 to 1.102 (Figure 5.19). This might correspond either to a higher input of fresh organic matter ($F^{14}C$ of ca. 1.15 in the 1990s) as a result of a positive impact of MAP on vegetation NPP or to a slight increase in the mineralization of the oldest compound of the topsoil, the two not being exclusive.

- **Impact of the Mean Annual Precipitation (MAP) on the $F^{14}C$ incorporation depth**

The MAP is a significant predictor for $F^{14}C$ incorporation (ϕ_3) with a posterior inclusion probability of 84%. The posterior mean estimation of the effect of the MAP **is 28 cm** (standard deviation of 10 cm) with a 99% probability of being positive.
Based on the results provided by the statistical model, we found that the MAP on the $F^{14}C$ incorporation (ϕ_3) is highly overestimated. It does not make sense for a 1 mm increment in the MAP to increase the depth corresponding to half $F^{14}C$ topsoil by 28 cm. Nevertheless, this can be interpreted as the injection of young carbon in depth.

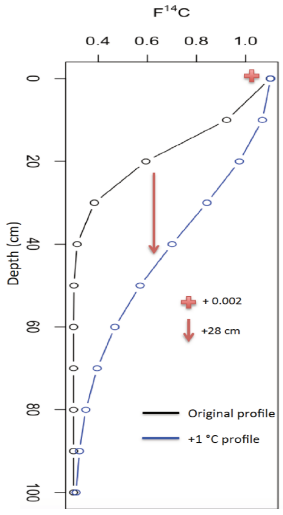


Figure 5.19: The impact of an increment of 1 mm in the Mean Annual Precipitation (MAP) on the topsoil $F^{14}C$ (ϕ_2) and the $F^{14}C$ incorporation (ϕ_3). The profile is plotted for ϕ_1 (deep $F^{14}C$) = 0.3, ϕ_2 (topsoil $F^{14}C$) = 1.1, ϕ_3 (the distance corresponds to half of the topsoil $F^{14}C$ amount) = 20 and $\phi_4 = 2$. The modified profile is shown in blue.

5.3.3.2 Impact of the Mean Annual Precipitation (MAP) on Soil carbon content

- **Impact of the Mean Annual Precipitation (MAP) on the Topsoil carbon content**

  There is a 52% chance that the **MAP will affect the topsoil carbon content** (ω_2). The posterior mean effect of MAP, on the topsoil carbon content, **is estimated to 11 kg/g** with 10 kg/g of standard error and 88% chance of being positive. Thus, a 1 mm increment in MAP, on average and over all vegetation and soil types, might result in increasing the topsoil carbon content by 11 kg/g. The combination of this result and the previous one obtained on the topsoil $F^{14}C$ tends to favor the assumption of a higher input of fresh organic carbon due to higher vegetation NPP. However, this does not exclude the second one, *i.e.* an enhanced degradation of the oldest components of the topsoil that can be compared to the concept of the priming effect (Fontaine et al., 2003). In fact, input in fresh carbon can result in an increase of soil carbon content, an apparent shift of topsoil $F^{14}C$ towards this new pool $F^{14}C$ and a priming effect resulting in the loss of old organic matter, thus maintaining a positive total budget of the carbon stock and enhancing the decrease of topsoil $F^{14}C$.

- **Impact of the Mean Annual Precipitation (MAP) on the incorporation depth**

  There is a 62% chance that the MAP affects the carbon content incorporation depth (ω_3). The mean posterior effect of MAP, on the soil carbon content incorporation, **is estimated at 76 cm**. As for $F^{14}C$, this value does not reflect what happens in reality and a very poor estimation is reported for the regression effects of the soil carbon content incorporation.

Based on the mathematical approach, Mean Annual Precipitation (MAP) is an influential predictor on soil carbon dynamics for both carbon content and $F^{14}C$. It leads to an increase in the soil carbon content associated to a tiny decrease of topsoil $F^{14}C$. This result can be related to both the dilution of topsoil components by the newly added compounds (increase of vegetation production as a result of enhanced MAP) and to the priming effect that results in a loss of old soil compounds. MAP also impacts the profile shape of both carbon content and $F^{14}C$ by increasing the depth of incorporation. The estimation of the effect is however not realistic.

5.3.4 Aridity Index (AI)

Reminder

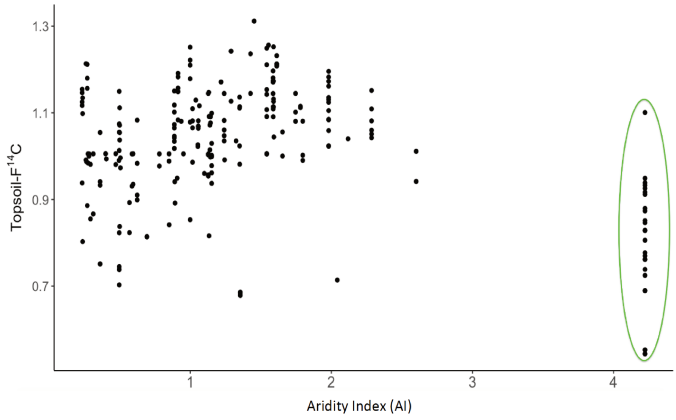
Aridity Index (AI) was selected as an influential predictor for topsoil (ϕ_2) and incorporation depth (ϕ_3) of the $F^{14}C$ profile and for none of the latent variables of the carbon content profile.

Aridity influences soil carbon inputs by affecting the production of above (loss of plant cover) and below ground plant biomass, water infiltration, microbial biomass and mineralization processes and thus the biogeochemical cycle of nutrients (Ren et al., 2018). A study done by Maestre et al. (2015) showed that the diversity and abundance of soil bacteria and fungi are reduced as aridity increases in global drylands. A recent study by Jones et al. (2018), to investigate the response of soil microbial communities to water and carbon availability across an aridity gradient (semi-arid, arid and hyper arid) within the Atacama, showed that even under hyper arid conditions, very low levels of microbial activity and carbon turnover occur. This result is in line with the one obtained by Rabbi et al. (2015) who showed that aridity has a strong negative influence on the soil organic carbon stock, based on a Structural Equation modeling applied to 1482 sites surveyed across the major agricultural region in Eastern Australia (AI varies from less than 0.07 (Eastern Australia coast) to more than 0.65 (moving away from the Eastern Australia coast)).

5.3.4.1 Impact of Aridity Index (AI) on topsoil $F^{14}C$

The Aridity Index (AI) is detected as influential for topsoil $F^{14}C$ with a posterior inclusion probability of 53%. The impact of the AI obtained by the posterior mean **is estimated at -0.029±0.022** with a 91% probability of being negative. **In that respect, the increase of 0.1 of AI leads to a** $F^{14}C$ **decrease of 0.0029**. In other words, the increase of 0.1 in AI will lead to a higher mean residence time of soil organic matter. Therefore, this result does not make sense since we expected that topsoil radiocarbon in the humid regions to be characterized by a higher $F^{14}C$ value (fresh carbon input) than in the arid regions.

To investigate this point further, we plotted the distribution of topsoil $F^{14}C$ (depth higher than 10 cm) according to the aridity index for all 131 profiles (Figure 5.20). This figure underlines a visually distinguishable increasing trend of the topsoil $F^{14}C$ with the increment in the AI up to a value of 3. In contrast, the green circle highlights a weird behavior of some points with a higher Aridity Index (AI = 4.223). A closer look at the database profiles shows that these points belong to the same 7 sites from Moor House Nature Reserve, that we already pointed out as not real mineral soils. They are further characterized by a very wet climate (MAP of 1665 mm) without efficient evaporation. They thus appear to undergo an equatorial monsoon climate whereas they are in a temperate region. This explains why the effect of the Aridity Index has a negative impact on the topsoil $F^{14}C$ instead of being positive.



Value	Climate Class
< 0.03	Hyper Arid
0.03 - 0.2	Arid
0.2 - 0.5	Semi-Arid
0.5 - 0.65	Dry sub-humid
> 0.65	Humid

Figure 5.20: Distribution of topsoil radiocarbon from the 131 database profiles where the depth does not exceed 10 cm according to the Aridity Index (AI). The boundaries that define various degrees of aridity and the approximate areas involved are given in the table on the right. The more AI tends to 0, the more arid the area is. In contrast an AI higher than 0.65 refers to a humid zone.

5.3.4.2 Impact of Aridity Index (AI) on $F^{14}C$ incorporation depth

The aridity index was detected as significant for the $F^{14}C$ incorporation depth with a posterior inclusion probability of 94%. The mean posterior effect is **estimated at -16 cm** with 9 cm of standard deviation (97% chance of having a negative effect). Thus, an increment of 0.1 in AI will decrease the $F^{14}C$ incorporation depth by 1.6 cm. Although more reasonable in extent, a physical interpretation of this parameter is not evident.

Aridity index is detected as significant only for the topsoil $F^{14}C$ and $F^{14}C$ incorporation depth. AI was expected to have a positive impact on topsoil $F^{14}C$ instead of having a negative impact. This result is due to 7 particular sites, all from the Moor House Nature Reserve in the United Kingdom that have a distinctive behaviour characterized by a very high AI and lower topsoil $F^{14}C$ measurements compared with the other sites in the database.

5.3.5 Seasonal temperature shift (Dif_T)

Reminder

Seasonal temperature shift (Dif_T) was selected as an influential predictor for topsoil (ϕ_2) and incorporation depth (ϕ_3) of the $F^{14}C$ profile and for the incorporation depth (ω_3) of the carbon content profile.

The greater the seasonal temperature shift, the greater the discrepancy between the summer and winter month temperatures. During the winter period characterized by a low temperature, which reduces the soil moisture due to freezing and removes water from the liquid phase, the decomposition rate of the soil organic matter slows down. In contrast, during the summer period characterized by a high temperature, microorganisms activity is favored and mobilizes the degradation of the soil organic matter protected during the winter period (Naganawa et al., 1989; Anderson, 1973). The question that remains is, in the long-term, what will be the most important seasons? which season will impose its fingerprint on the soil carbon dynamics? Very few studies deal with that aspect in the literature.

5.3.5.1 Impact of seasonal temperature shift (Dif_T) on topsoil $F^{14}C$

The seasonal temperature shift was detected as an influential predictor for the topsoil $F^{14}C$ with a posterior inclusion probability of 60%. The posterior effect regression of Dif_T **is estimated at -0.037** (standard deviation of 0.01) with a 99% probability of having a negative effect. For example, **an increment of 1°C in seasonal shift between the warmest and coldest months, for a radiocarbon topsoil profile of 1.1, will increase the mean** ^{14}C **age of soil organic carbon**, *i.e.* contributing to a higher mean residence time.

Except for the profiles from the same English site in a peat-derived environment (green circle in Figure 5.21), a closer look at the database also shows a general trend of topsoil (less than 10 cm depth) $F^{14}C$ decreasing while continentality, i.e. Dif_T, increases. It is thus likely that the evaluation of the impact of this predictor and more likely of the associated deviation are biased by the presence of these 7 sites. A clearer signal might be obtained from a second evaluation of the database excluding these 7 profiles.

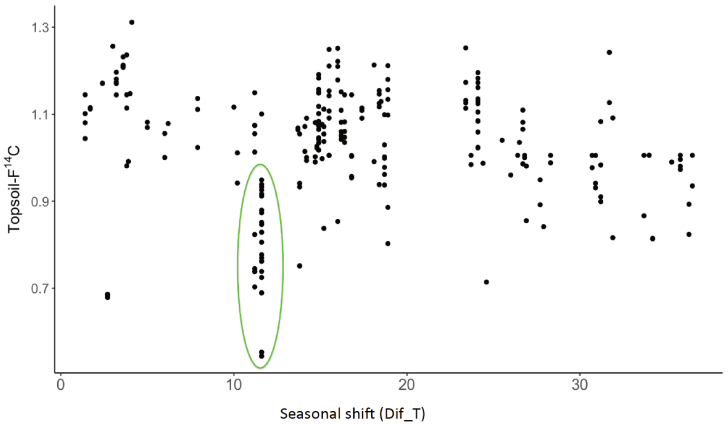


Figure 5.21: Distribution of the topsoil (less than 10 cm depth) $F^{14}C$ from the database profiles versus the seasonal shift (temperature difference between the hottest and the coldest month of the year). The green circle highlights the specific UK sites from the Moor House Nature Reserve "British profiles".

5.3.5.2 Impact of seasonal temperature shift (Dif_T) on $F^{14}C$ incorporation depth

The seasonal temperature shift was detected as influential for the $F^{14}C$ **incorporation depth** with a posterior inclusion probability of 72%. The posterior mean effect **is estimated at 6 cm** with 7cm of standard deviation (83% chance of having a positive effect). Thus, an increment of the shift between the warmest and coldest months of 1°C increases the depth incorporation, which corresponds to half of the topsoil radiocarbon activity, by 6 cm.

5.3.5.3 Impact of seasonal temperature shift (Dif_T) on carbon content incorporation depth

The seasonal temperature shift (Dif_T) was detected as significant for carbon content incorporation depth (ω_3) with a posterior inclusion probability of 68%. The posterior mean effect **is estimated at -9 cm** with 5cm of standard deviation and a 96% chance of having a negative effect.

Our mathematical approach provides decision-making elements in the debate on the impact of the temperature seasonality on soil organic carbon dynamics. Even if evaluation of the impact extent is hampered by the British profiles, there is a trend towards a positive impact on the mean residence time: the higher the continentality, the higher the residence time.

5.3.6 Minimum monthly precipitation (min_P)

Reminder

The minimum monthly precipitation (min_P) was selected as an influential predictor only for incorporation depth (ϕ_3) of the $F^{14}C$ profile and for none of the latent variables of the carbon content profile.

5.3.6.1 Impact of the Minimum monthly precipitation (min_P) on $F^{14}C$ incorporation depth

The minimum monthly temperature was detected as influential for the $F^{14}C$ incorporation depth with a posterior inclusion probability of 55%. The posterior mean effect **is estimated at -4cm** with 7cm of standard deviation (70% chance of having a negative effect). Thus an increment of 1 mm in the minimum monthly precipitation decreases the $F^{14}C$ incorporation depth by 4 cm. At a constant level, $F^{14}C$ thus decreases with the temperature increases, reflecting the loss of ^{14}C enriched components, maybe some decades old compounds whose carbon derives from the nuclear bomb peak.

Soil radiocarbon and soil carbon content incorporation (ϕ_3 and ω_3)

The estimation of regression effects corresponding to ϕ_3 (the $F^{14}C$ incorporation depth) and the ω_3 (the soil carbon content incorporation depth) indicates a poor estimation of what can happen in reality. This bad estimation can be explained by the non linear link between these latent variables, ϕ_3 and ω_3 , and the responses of $F^{14}C$ and soil carbon content respectively.

5.3.7 Soil type (Soil)

Recall

Soil type was identified as influential for all latent variables for both $F^{14}C$ profile and carbon content profiles

Many soils are marked by climate and type of vegetation (Legros, 2007). For example, Gleysols and Podzols are characteristic of cold regions. Cambisols, Luvisols and Podzols are conditioned by temperate climates (Spaargaren, 2001). Ferralsols and Plinthosols are tropical soils with forest cover. In addition, Chernozem, Kastanozems and Phaeozems are associated to steppe and grassland vegetation cover (http://www.isric.org), under climatic regimes that range from cool temperate to warm Mediterranean (https://www.britannica.com). Various soil physical and chemical properties such as the clay content are reported to control the organic matter decomposition rates (Balesdent et al., 2000). The clay and silt content is assumed to be positively correlated with the soil organic carbon (Paul et al., 2008).

5.3.7.1 Impact of soil type on mean residence time ($F^{14}C$)

- **Impact of soil type on deep $F^{14}C$**

  Jobbágy and Jackson (2000) underline that 56% of soil carbon globally can be found below 1 meter. **The soil type was detected as influential for the deep radiocarbon response with posterior inclusion probabilities of 51%**. The fusion of the levels by soil type obtained by the BEF (with k = 50) discriminates between the levels (see colors in Figure 5.22 and major lines in Table 5.16) and proposes a clustering that outputs 4 different groups.

  Recap of the result of levels fusion for soil type for the deep $F^{14}C$

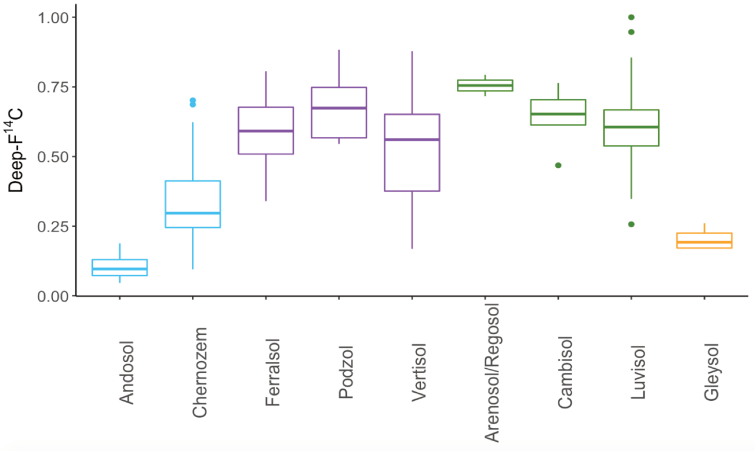


Figure 5.22: Distribution of deep $F^{14}C$ (at least 100 cm of depth) according to the soil type. The boxplots of soil categories having the same color belong to the same cluster.

The BEF approach to deep soil $F^{14}C$ roughly follows the distribution of deep soil $F^{14}C$ from the database (Figure 5.22 and last column of Table 5.16). Indeed, clusters 1 (blue in Figure 5.22) and 4 (orange) are associated to the lowest recorded $F^{14}C$ whereas clusters 2 (purple) and 3 (green) are associated to the highest values. The distinction between clusters 2 and 3 can be justified by the range of the associated quantiles. However, the grouping of Andosol and Chernozem (blue), or the exclusion of Gleysol (orange), cannot be explained only by the $F^{14}C$ values in themselves.

The BEF approach to deep soil $F^{14}C$ rather reflects the mineralogical properties of deep horizons, as already mentioned by Mathieu et al. (2015). It is, indeed, interesting to confront the fusion result with the clay property of the soil type (Table 5.16).
Thus, BEF combines Chernozem and Andosol (cluster 1, blue boxplots in Figure 5.22) both of them being rich in clay and particularly in high activity clay (Table 5.16). However, according to clay amount (see Table 5.16), Vertisol should have been classified within the same group. Vertisol and Chernozem both contain high-activity smectite clays in combination with exchangeable calcium saturation, that contributes to stabilizing carbon (Lützow et al., 2006). Moreover, as well as Chernozem and Andosol that stabilize soil organic matter and consequently show low $F^{14}C$, Vertisol should also have shown low $F^{14}C$ and that is not the case (Figure 5.22 and Table 5.16). The small number of Vertisol profiles, only 5 with data below 100cm, may be the reason for the misattribution of this type of soil.

The second cluster groups Podzol and Ferralsol (cluster 2, purple boxplot in Figure 5.22) (if we exclude Vertisol). This cluster of soil types is identified by a low activity clay, in varying amount (Table 5.16). Here also, based on the mineralogical properties, Vertisol should not have been merged in this cluster.

The third cluster (green boxplots in Figure 5.22) contains the Arenosol/Regosol group, the Cambisol group and Luvisol. The clay amount recorded for Arenosol/Regosol in Table 5.16 corresponds only to the Arenosol soil type. This cluster is characterized by a low amount of clay, of rather medium activity type.

The final cluster (orange boxplot in Figure 5.22) corresponds to Gleysol only which is characterized by an undefined amount of clay. It is not clear why Gleysol has not been merged in another cluster: cluster 1 with which it shares the low value of $F^{14}C$, cluster 3 with which it shares the low activity type of clay.

Cluster	Merged WRB group	WRB soil type	Amount of clay	Type of clay	Median deep soil $F^{14}C$	[25% q; 75% q]
1 (blue)	Andosol (12)	Andosol (12)	High	Very high activity	0.09	[0.07;0.13]
	Chernozem (19)	Chernozem (16) Kastanozem (1) Phaeozem (2)	Medium	High activity	0.29	[0.24;0.42]
2 (purple)	Vertisol (7)	Vertisol (7)	High	High activity	0.56	[0.37;0.65]
	Podzol (16)	Podzol (16)	Very Low	Low activity	0.67	[0.57;0.75]
	Ferralsol (18)	Ferralsol (14) Nitisol (4)	Medium to high	Low activity	0.59	[0.51;0.68]
3 (green)	Arenosol/Regosol (7)	Arenosol (3) Regosol-Arenosol(1) Leptosol (3)	Low	Very low activity	0.75	[;0.73, 0.77]
	Cambisol (15)	Cambisol (15) Fluvisol (1)	Low	medium activity	0.65	[0.61;0.70]
	Luvisol (27)	Luvisol (27)	Low	Medium activity	0.60	[0.54;0.67]
4 (orange)	Gleysol	Gleysol (7) Planasol (1) Plinthosol (1)	Undefined	Low activity	0.19	[0.17;0.22]

Table 5.16: Amount and type of clay generally observed in the soil types from the database. Types of soil are ranked according to the result of Bayesian Effect Fusion of soil type levels for deep soil radiocarbon activity (ϕ_1). Cluster number and color are the ones used in Figure 5.22. The number of profiles associated to each type of soil and each group of type of soils are provided in brackets. Column six refers to the median value of deep (higher than 100 cm deep) soil $F^{14}C$ from the database (line inside rectangle in Figure 5.22). The last column refers to 25% and 75% quantiles (q= quantile) (the upper and lower rectangle bounds in Figure 5.22) * The result of the clay amount and the type of clay corresponds to Arenosol soil type, only.

- **Impact of soil type on topsoil $F^{14}C$**

According to the posterior inclusion probabilities, **the soil type has a higher posterior probability of affecting the topsoil $F^{14}C$ (100%) than the deep one (51%)**. The fusion of the soil type levels for topsoil $F^{14}C$ proposes 5 different groups. The clusters for topsoil are completely different from those for deep soils.

Recap of the result of levels fusion by soil type for the topsoil radiocarbon

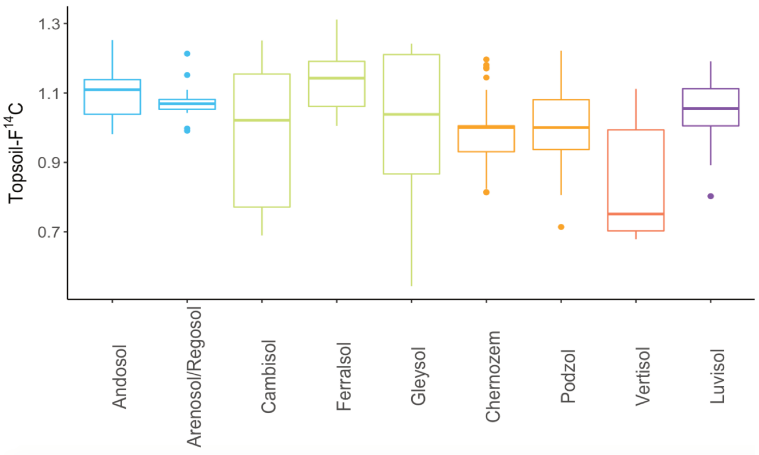


Figure 5.23: Distribution of the topsoil $F^{14}C$ (less than 10 cm of depth) according to the soil type. The boxplots of soil categories having the same color belong to the same cluster.

Based on $F^{14}C$ values from the database, clusters 1 and 2 on the one hand and 3 and 5 on the other hand show a similar $F^{14}C$ distribution around 1.09 and 1 respectively, whereas the fourth cluster shows quite different behavior with $F^{14}C$ around 0.75.

The first cluster grouping Andosol and Arenosol/Regosol (Figure 5.23 - blue boxplots) characterizes soils conditioned by parent material: Andosol rapidly develops on weathering pyroclastic deposits and on volcanic material and Arenosol is a sandy soil. The second cluster is formed from Cambisol, Ferralsol and Gleysol (Figure 5.23, green boxplots). Cambisols in the humid tropics are predominant at medium altitudes in hilly and mountain regions but also in deposition areas and in eroding lands at lower altitude where they occur alongside genetically mature residual soils (e.g. Acrisols or Ferralsols). Likewise Ferralsol mostly develops in tropical climate zones on old parental material. Gleysol shows two geographical distributions: polar regions and tropical and subtropical regions. It seems then that clustering mostly accounts, here, for gleysol from tropical zones.

Chernozem and Podzol are associated in a third cluster (Figure 5.23, orange boxplots). They share the properties of having a characteristic surface layer rich in humus and being mostly associated to grass vegetation. In addition, both Chernozem and Podzols are found in cold regions.

Vertisol constitutes a single group, the fourth cluster (Figure 5.23, red boxplot) characterized by cracking clay soils extending downward from the land surface and by evidence of strong vertical mixing of soil particles over successive wet and dry periods.

The last cluster (Figure 5.23, purple boxplot) contains Luvisol only. Luvisol is distinguished from the other soil types by having a subsurface horizon which has a distinctly higher clay content than the overlying horizon. Added to that, Luvisols are distinguished by a temperate climate under grassland vegetation.

It seems that the clustering operated by BEF to express the topsoil $F^{14}C$ latent variable illustrates the

textural quality of the soil profile and the difference between the surface layer and the remaining part of the profile.

- **Impact of soil type on $F^{14}C$ incorporation depth**

The soil type was detected as significant for $F^{14}C$ incorporation depth with 97% of posterior inclusion probability. The Bayesian Effect fusion identified 6 different clusters of soil types (Figure 5.24). However, we do not have a clear idea about the criteria that explain why the soil types were split into 6 different clusters.

Recap of the result of levels fusion by soil type for $F^{14}C$ incorporation depth

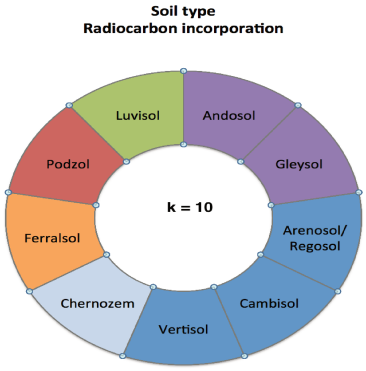


Figure 5.24: The fusion of levels for soil type for the $F^{14}C$ incorporation depth (ϕ_3) according to the Bayesian Effect Method with k = 10. Pie slices of soil categories having the same color belong to the same cluster.

The result of fusion of soil types for both topsoil $F^{14}C$ and deep $F^{14}C$ layers underlines that the $F^{14}C$ is more dominated by the climate/vegetation and soil texture at the topsoil and by clay content for deeper layers.

5.3.7.2 Impact of soil type on carbon content

The soil type was detected as significant for all latent variables, ω_1, ω_2 and ω_3, of the soil carbon content profile by the Bayesian Effect Fusion selection method. However, based on Figure 5.25 and Figure 5.26, no clear physical interpretations can be proposed for the different clusters of fused soil types identified.

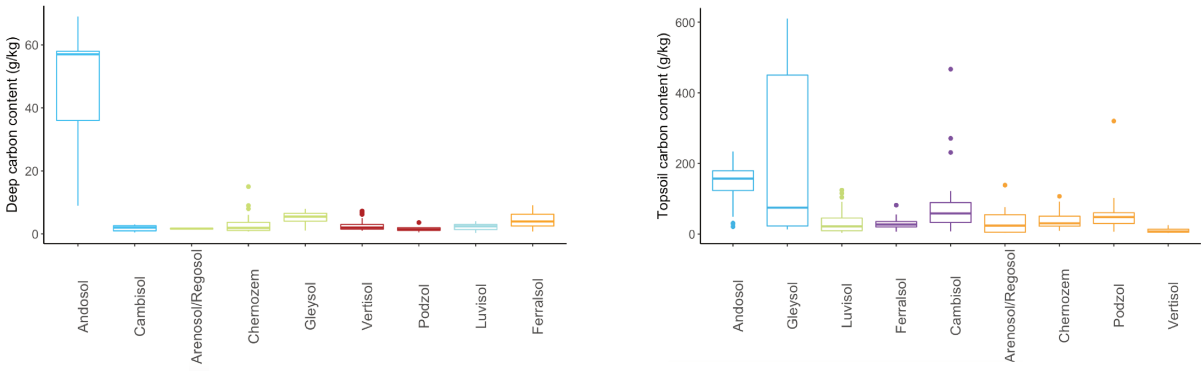


Figure 5.25: Distribution of both topsoil (panel a) and deep soil carbon content (panel b) (depth $\leq$ 10 cm and depth $>$ 100 cm) according to the soil type is illustrated by boxplots. The boxplots of soil categories having the same color belong to the same cluster.

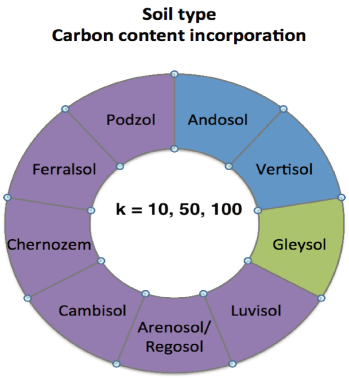


Figure 5.26: The results of the soil type levels fusion for the carbon content incorporation depth (ω_3)) obtained by the BEF. The same fusion of levels was identified under k = 10, 50 and 100. Pie slices of soil categories having the same color belong to the same cluster.

We have not identified physical interpretations for the results of fusion of ecosystem types for both topsoil and deep soil carbon content layers.

5.3.8 Ecosystem type (land)

Reminder

Ecosystem was identified as influential for all latent variables for both $F^{14}C$ profile and carbon content profile

5.3.8.1 Impact of ecosystem type on mean residence time ($F^{14}C$)

- **Impact of ecosystem type on topsoil and deep $F^{14}C$**

Summing vegetation and land use, ecosystem type was detected as significant for both topsoil and deep $F^{14}C$ with the following posterior inclusion probabilities: 53% and 56%. In both cases, the levels fusion of ecosystem type singles out only two clusters: field as cluster 1 and the remaining ecosystem types for cluster 2 for the deep soil; cultivated grassland as cluster 1 and the remaining ones as cluster 2 for top soil (Figure 5.27).

There is no clear apparent distinction between the two clusters based on $F^{14}C$ from the database. For deep soil, one can guess a trend toward higher $F^{14}C$ values associated to cluster 1 (field) by comparison with cluster 2 (all other types of ecosystem) (Figure 5.27, panel b). However, it is well known that cultivation impacts soil organic matter by accelerating soil organic carbon mineralization and reducing the input of fresh plant–derived organic matter to all layers. This results in a reduced mean residence time of soil compounds and reduced renewal of the organic pool in all horizons. This might be the reason why "field" is considered differently from other ecosystem types for deep soil but does not explain why it was not distinguished for the topsoil.
We could not find any clear reason why cultivated grassland (mostly pasture) appears in a separate cluster for topsoil.

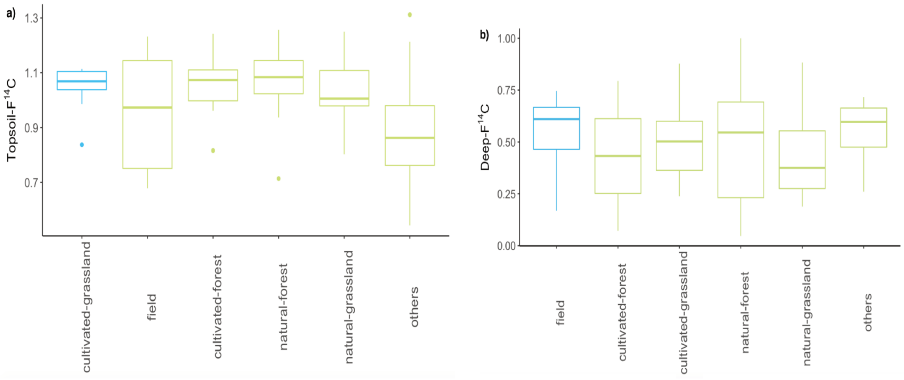


Figure 5.27: Distribution of both topsoil $F^{14}C$ (panel a) and deep $F^{14}C$ (panel b) (depth lower than 10 cm and depth higher than 100 cm, respectively) according to the ecosystem type. The boxplots of soil categories having the same color belong to the same cluster.

- **Impact of ecosystem type on $F^{14}C$ incorporation depth**

The ecosystem type was detected as significant with the Bayesian Effect fusion since not all its levels are merged. However, it is difficult to interpret the result of levels fusion for soil type according to the posterior fusion probability (Figure 5.28). Here, we cannot say that the fusion result distinguished the natural ecosystem from the cultivated one, since field belongs to the cluster where natural ecosystems dominate.

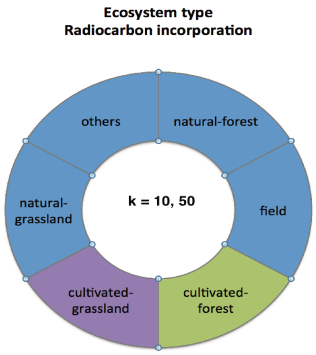


Figure 5.28: The fusion of levels of ecosystem type for the $F^{14}C$ incorporation depth (ϕ_3). The slices of ecosystem categories having the same color belong to the same cluster. The fusion of ecosystem levels is the same for k = 10 and 50.

5.3.8.2 Impact of ecosystem type on soil carbon content

- **Impact of ecosystem type on topsoil and deep soil carbon content**

Ecosystem type was detected as significant for both topsoil and deep carbon content. The levels fusion of ecosystem types highlights four clusters for both top and deep soil (Figure 5.29). For topsoil, the clusters are as follows: 1- field and "others", 2- natural forest, 3- natural grassland, 4- cultivated forest and cultivated grassland. For Deep soil, the clusters are as follows: 1- field, 2- "others" and natural forest, 3- natural grassland, 4- cultivated forest and cultivated grassland. The only difference is that the category ”others”, grouping 13 profiles of the database, is merged with natural forest for the deep soil carbon content, and with cropland for the topsoil carbon content.

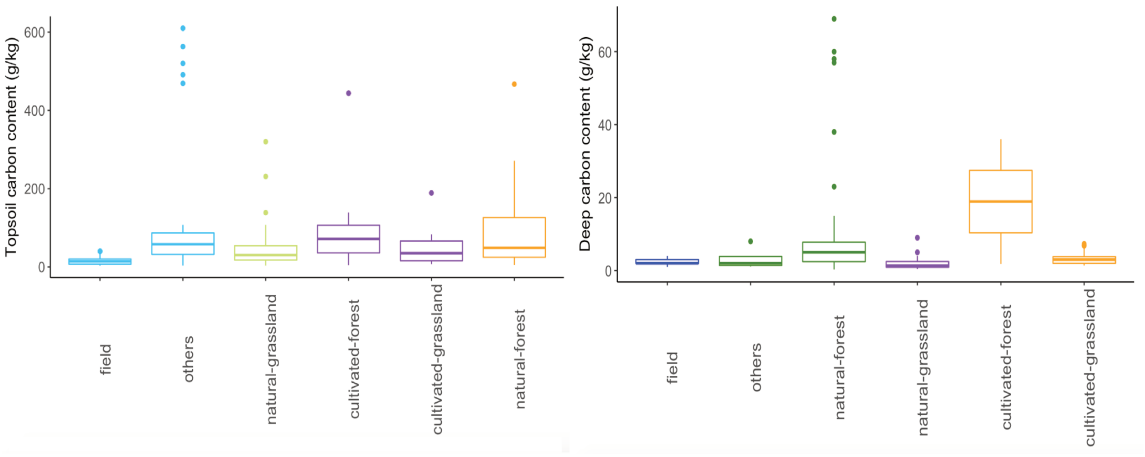


Figure 5.29: Distribution of both topsoil (panel a) and deep soil carbon content (panel b) (depth $\leq$ 10 cm and depth $>$ 100 cm) according to the ecosystem type is illustrated by boxplots. The boxplots of soil categories having the same color belong to the same cluster

The soil carbon content varies among the different types of vegetation cover (forest, grassland, etc.) and the land use practices (natural or cultivated). For example, forest tends to have the largest input of vegetation carbon. The average Net Primary Production (NPP) varies from 2200 tons per km^2 /yr for tropical rainforest to 800 tons per km^2 /yr for boreal forest (Jackson et al., 1997). Tropical rainforest and boreal forest occupy respectively, 11% and 22% of the percentage of earth's land surface. In addition, forests are characterized by the most recalcitrant carbon and a lower decomposition rate especially for forest with long roots. The smallest input of NPP corresponds to cropland where the carbon inputs are among the most labile as a result of biomass removal in the harvest products and the land use practices than can open aggregates to weathering and microbial breakdown. Soil under forest has significantly higher values of soil organic carbon (59.35 Mg/ha) than pasture (42.48 Mg/ha) and arable land (23.63 Mg ha1) (Ali et al., 2017).

Here, we can interpret the classification according to the different Net Primary production (NPP) between forest, grassland and cropland. Furthermore, the classification of cultivated grassland and forest together may highlight the effect of land use practices on soil carbon content dynamics. Cultivation may change the ground surface and root distribution of plants. Furthermore, cultivation has a potential to destroy soil structure and make soils more prone to other forms of degradation, such as erosion.

Nonetheless, the signal is not that clear and we may wonder if another type of vegetation and land-use grouping would not have been more appropriate. For example, keeping only 3 major groups: cropland, grassland and forest, keeping the notion of man induced landscape only for cropland.

- **Impact of the ecosystem type on carbon content incorporation depth**

The ecosystem type is detected as influential for carbon content incorporation depth (ω_3) by the Bayesian Effect Fusion. However, the result of fusion of levels for ecosystem type discriminates natural forest from all other ecosystem types. The result of fusion does not have any physical interpretation (Figure 5.30).

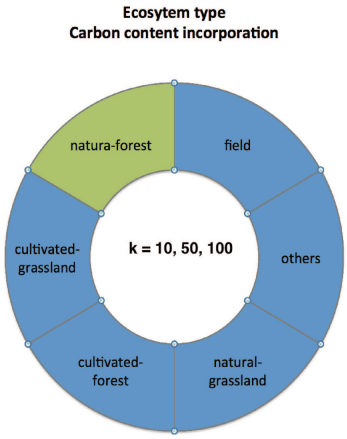
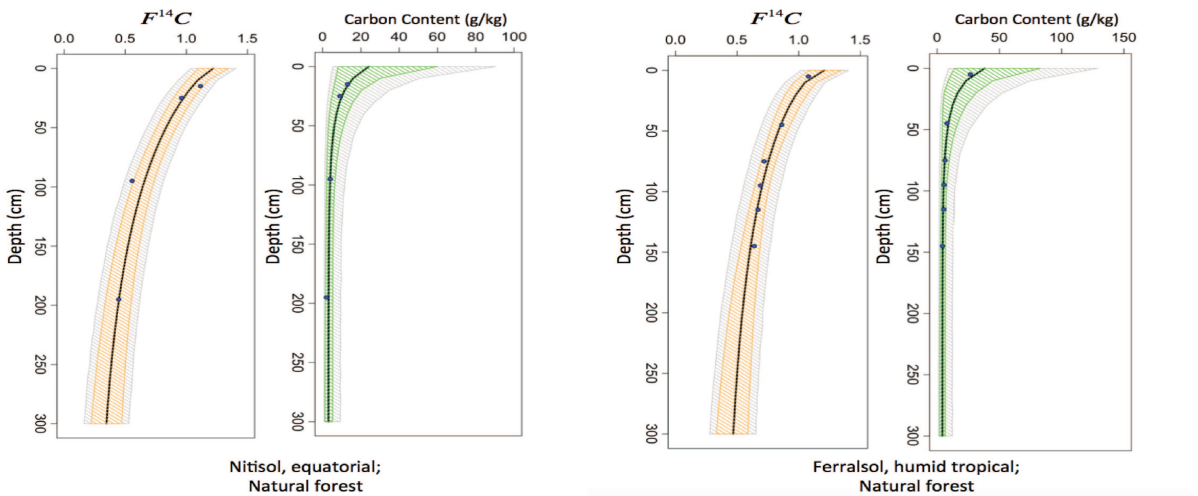
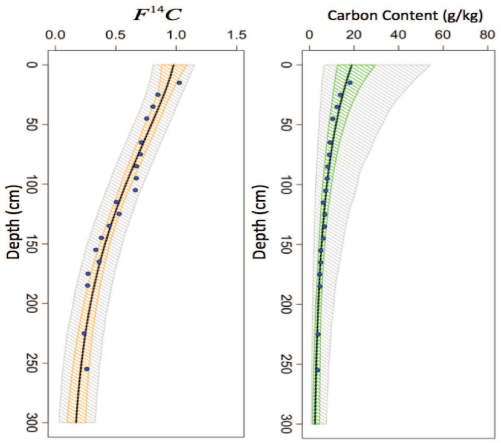


Figure 5.30: The fusion of levels of ecosystem type for the carbon content incorporation depth (ω_3). The slices of ecosystem categories having the same color belong to the same cluster. The same levels fusion of ecosystem type is obtained under different values of k : k = 10, 50 and 100.

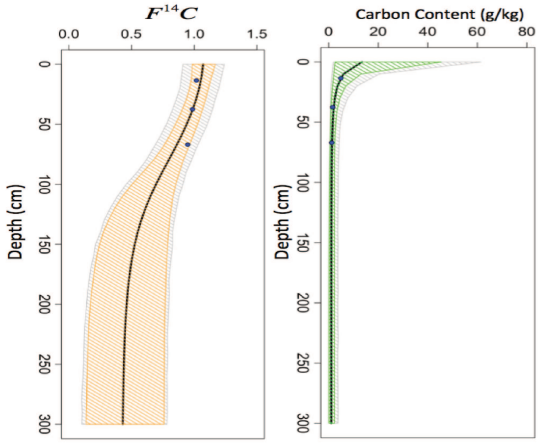
5.4 Synthetic representation of the dependence of $F^{14}C$ and soil carbon content on soil-climate-biomes

In order to compare the predicted profiles with observed $F^{14}C$ and soil carbon content of the major world soil groups present in contrasting climates, we chose ten representative soil profiles from the database (Figure 5.31). The predicted profiles are obtained according to the best sub-models identified for the soil radiocarbon and soil carbon content dynamics respectively. Figure 5.31 illustrates the overall dependence of $F^{14}C$ and soil carbon content profiles on environmental variables in a holistic approach that implicitly accounts for the interacting roles of all variables on radiocarbon dynamics.

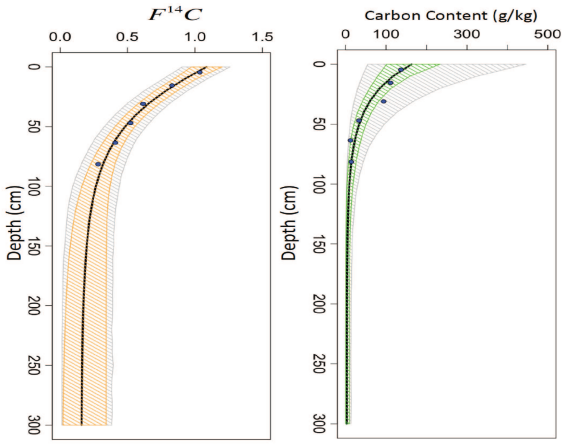




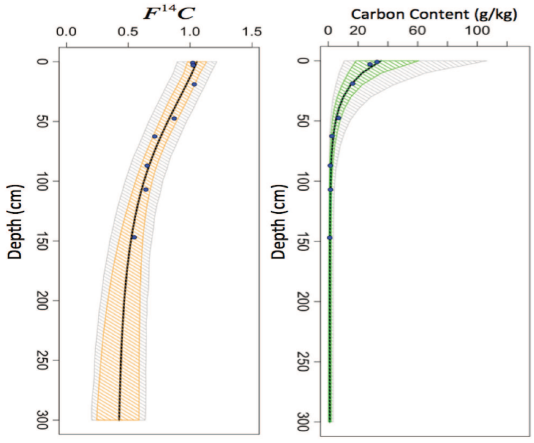
Vertisol, dry tropical;
Cultivated grassland



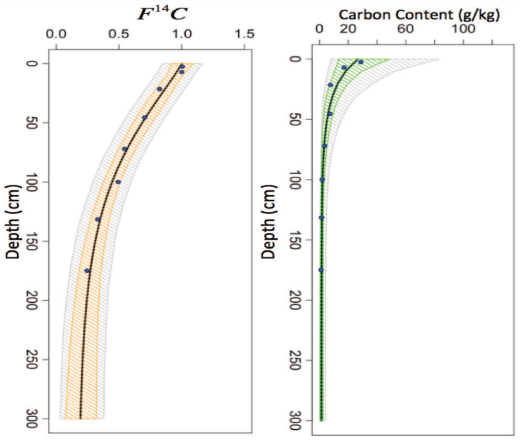
Luvisol, dry mediterranean;
Natural grassland



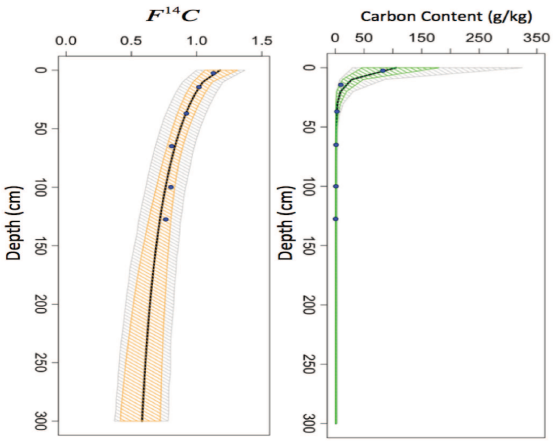
Andosol, humid cool;
Natural forest



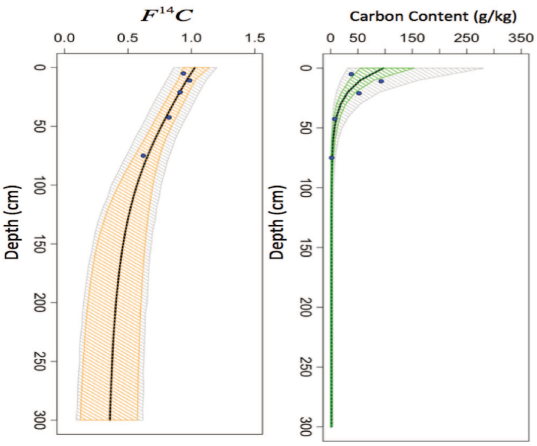
Luvisol, humid cool;
field



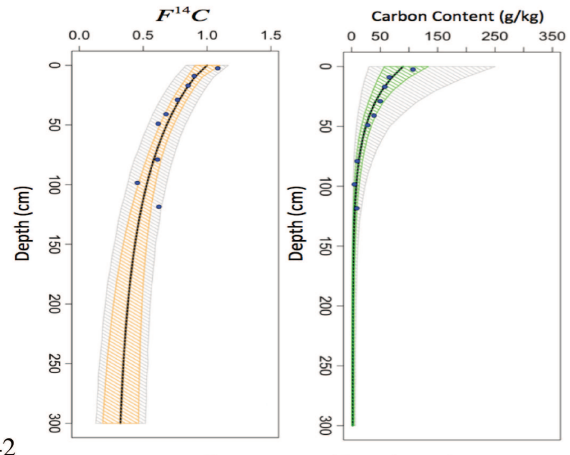
Chernozem, dry continental;
Natural grassland



Cambisol, humid cool;
Natural forest



Podzol, humid cool;
Natural forest



Chernozem, cold continental;
Natural grassland

Soil group	Nitisol	Ferralsol	Vertisol	Luvisol	Andosol
Vegetation	natural forest	natural forest	cleared grassland	natural grassland	natural forest
MAT/Dif_T(°C)	26.4/1.4	20.5/6.2	19.5/14.3	16/18.4	10.9/26.5
MAP(mm)	2237	1270	666	369	1113
Aridity Index	1.75	1.00	0.40	0.23	1.33
min_P (mm)	25	25.4	31.9	0.8	44.6
sampling year	1992	1994	1997	1978	2001
Reference data	Pessenda et al., 98	Pessenda et al., 96	Krull et al., 03	Baisden et al., 02	Katsuno et al., 10
Country	Brazil	Brazil	Australia	USA	Japan

Soil group	Luvisol	Chernozem	Cambisol	Podzol	Chernozem
Vegetation	field	natural forest	cleared grassland	field	natural forest
MAT/amplitude(°C)	9.9/14.8	9.4/26.8	8.4/18.5	6.4/18.7	5.5/31.2
MAP(mm)	698	382	673	723	507
Aridity Index	1.05	0.28	1.04	1.15	0.62
min_P (mm)	48.1	7.6	41.9	45.2	28.5
Sampling year	2009	1994	1996	1996	1997
Reference data	Jagercikova et al., 14	Leavitt et al., 07	Rumpel et al., 02	Schulze et al, 09	Torn et al., 02
Country	France	USA	Germany	Germany	Russia

Figure 5.31: Synthetic view of the dependence of soil $F^{14}C$ and carbon content on soil–climate–biome. Ten sites from the database were selected as representative of 10 major biomes, taking into account only the explanatory covariates detected as significant for soil $F^{14}C$ and soil carbon dynamics respectively. The orange (the green) band corresponds to the confidence in the local (within site) estimate of $F^{14}C$ (soil carbon content), and the gray band corresponds to the between-site variability of soil with similar environmental variables. Dark lines represent the sampled horizon of observed data and the blue points the real $F^{14}C$ (soil carbon content) measurements.

Briefly, it is interesting to note that simulated profiles are very close to measured data, for both carbon content and $F^{14}C$ profiles. Simulation reflects the general shape of the original profile but misses some specific features, as is expected from a model. The mean estimated topsoil $F^{14}C$ ranges between 0.97 and 1.22 and the mean estimated deep radiocarbon varies between 0.10 and 0.45 in all ten soils. Generally speaking and excluding the Andosol-type soils, the surface $F^{14}C$ was the highest in the warm tropical climates and was the lowest in the coldest climates. The mean estimated carbon content ranges between 3.88 g/kg and 294 g/kg for topsoil and between 0.02 g/kg and 5.56g/kg for the deep soil. Soil carbon content estimation at the top of the profiles is associated to a very large variability that reflects the natural worldwide variability of carbon content (Tifafi et al., 2018). It appears thus that even by forcing the system by specifying soil type, climate and land-use, the variability remains very high. The variability of the simulated profile decreases with depth.

It is worth noting the specific fate of some profiles. For example, the "Vertic" fate of the Vertisol that shows a specific $F^{14}C$ profile with a very deep $F^{14}C$ incorporation depth, associated, as shown by the carbon content profile, with deep incorporation of fresh plant-derived organic material. Chernozem and boreal Podzol are deep soils with a long history, close to the fate of permafrost. This is also reflected by the very low $F^{14}C$ in depth. In contrary Ferralsol and Cambisol are young soils with rapid turnover and thus associated to a higher $F^{14}C$ in depth. Luvisol, also, shows a specific structure. It exhibits good drainage and this is clearly expressed in the carbon content profile of the Mediterranean Luvisol where a drastic decrease in carbon content occurs at the upper levels of the profile.

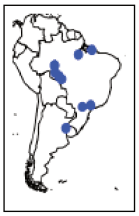
5.5 Predictive model applications in a context of global changes

5.5.1 Impact of a land use change

5.5.1.1 Conversion of equatorial forest to cropland impacts both topsoil carbon content and deep carbon dynamics

A study done by Noojipady et al. (2017) based on the satellite data on cropland expansion, forest cover and vegetation carbon stocks in the Cerrado biome showed that 29% of the carbon emission which is equivalent to 16.28 tg C yr^{-1} (tg = teragram = 10^{12} grams), between 2003 and 2013, is due to the conversion of forest to cropland.

What changes will be observed on the soil radiocarbon dynamic for a tropical forest profile?



To address this question, we focus on the Brazilian forest. We selected the nine profiles in the database that met the following criteria: Brazilian profiles associated to "natural forest" ecosystem type (Table 5.17). Three types of forests are represented: rainforest, "natural forest" (without further description in the original articles) and transition vegetation between natural forest and cerrado. Three Köppen-Geiger climate subgroups are represented: Equatorial monsoon (Am), Equatorial savannah with dry summer (As) and Temperate fully humid (Cf). Four soil types were reported for these profiles: four Ferralsol, two Nitisol, two Cambisol and one Podzol. To make the experimental design more balanced, please remember that Ferralsol and Nitisol were fused into Ferralsol.

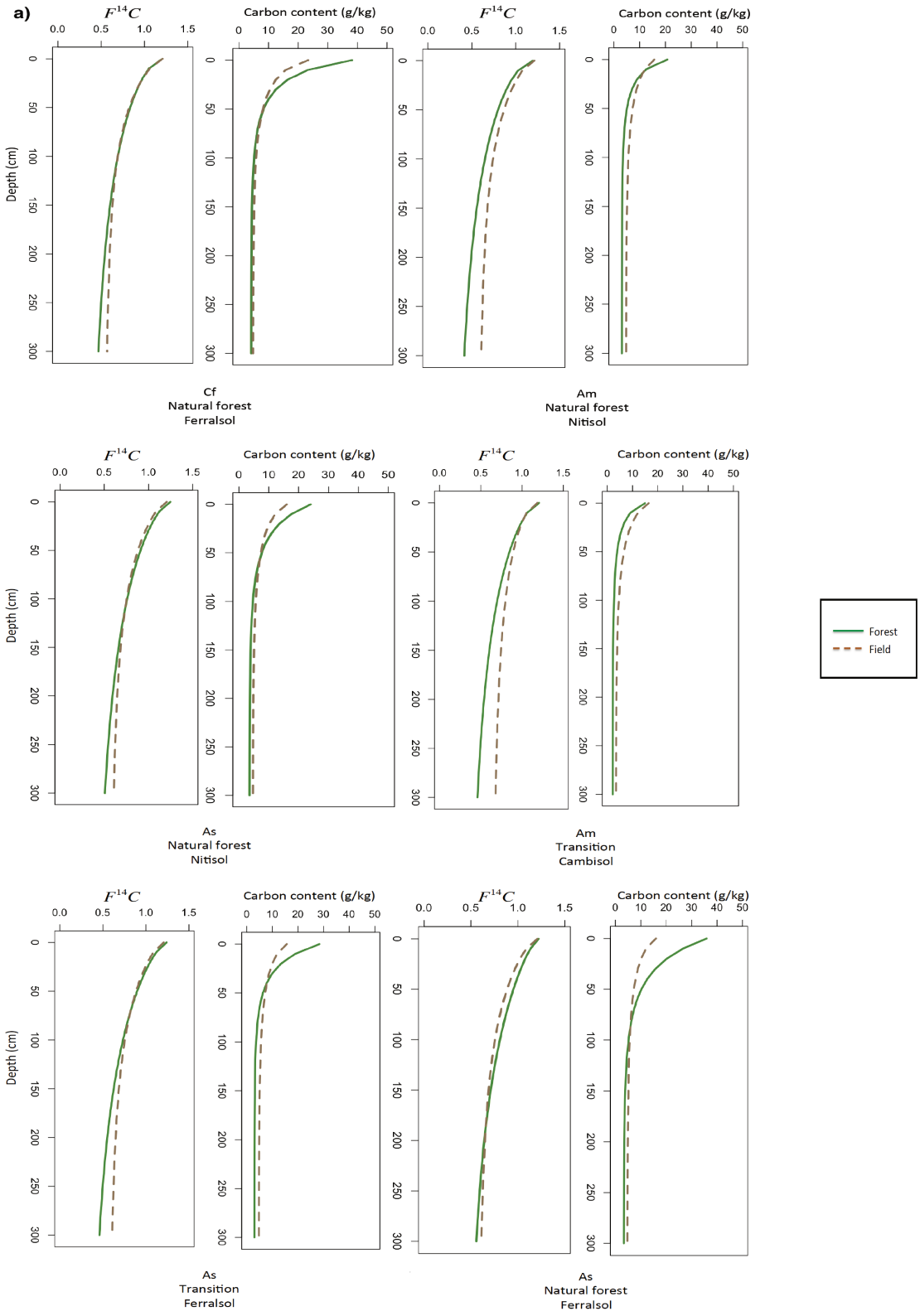
Soil group Vegetation	Ferralsol natural forest	Nitisol natural forest	Ferralsol transition	Ferralsol natural forest
MAT/Dif_T($°C$)	20.5/6.2	26.3/1.4	26.2/3.8	26.2/3.8
MAP(mm)	1270	2040	2004	1986
Aridity Index	0.99	1.59	1.43	1.43
min_P (mm)	25.4	42.7	10.8	10.8
Sampling year	1994	1992	1992	1992
Reference data	Pessenda et al., 96	Pessenda et al., 96	Pessenda et al., 98	Pessenda et al., 98
Köppen-Geiger climate	Cf	Am	As	As

Soil group Vegetation	Nitisol natural forest	Cambisol transition	Cambisol natural forest	Podzol rainforest
MAT/Dif_T ($°C$)	26/3	26.4/1.4	26.4/1.4	26.4/1.7
MAP(mm)	2040	2237	2237	2467
Aridity Index	1.55	1.75	1.75	1.78
min_P (mm)	14.4	25	25	43.4
Sampling year	1992	1992	1992	1988
Reference data	Pessenda et al., 98	Pessenda et al., 98	Pessenda et al., 98	Desjardins et al., 94
Köppen-Geiger climate	As	Am	Am	Am

Table 5.17: Climatic and environmental conditions for nine $F^{14}C$ Brazilian profiles under a "natural forest" from the database. Vegetation cover is reported according to the authors descriptions, "transition" is for the vegetation type at the transition between cerrado and natural forest. Köppen-Geiger subgroup is calculated according to Kottek et al. (2006) rule (see Chapter 2")

A posterior mean estimation profile was obtained from these nine profiles for both $F^{14}C$ and carbon content (green line in Figure 5.32). This mean profile can be roughly viewed as the average humid tropical profile under a natural forest, characterized by Mean Annual Temperature (MAT) of 25.6°C, Mean Annual Precipitation (MAP) of 2057 mm and Aridity Index of 1.56.

To predict the impact of land use conversion, in this case, to predict deforestation, each set of variables was associated to the nine profiles, only the ecosystem type was changed from natural forest to field. Climate and type of soil were kept constant. Using the best sub-model selected according to the best sub-models identified for $F^{14}C$ and carbon content, we predicted the nine profiles that would correspond to $F^{14}C$ and carbon content profiles in the case of "field" as ecosystem type. A mean profile was estimated from these nine converted profiles (brown dashed line in Figure 5.32). Comparison between the original and the converted mean profile helps in figuring out what can happen in the event of a conversion from forest to field, in a rather equatorial climate context.


a)
$F^{14}C$
Carbon content (g/kg)
Depth (cm)
Cf
Natural forest
Ferralsol
Am
Natural forest
Nitisol
As
Natural forest
Nitisol
Am
Transition
Cambisol
As
Transition
Ferralsol
As
Natural forest
Ferralsol
Forest
Field

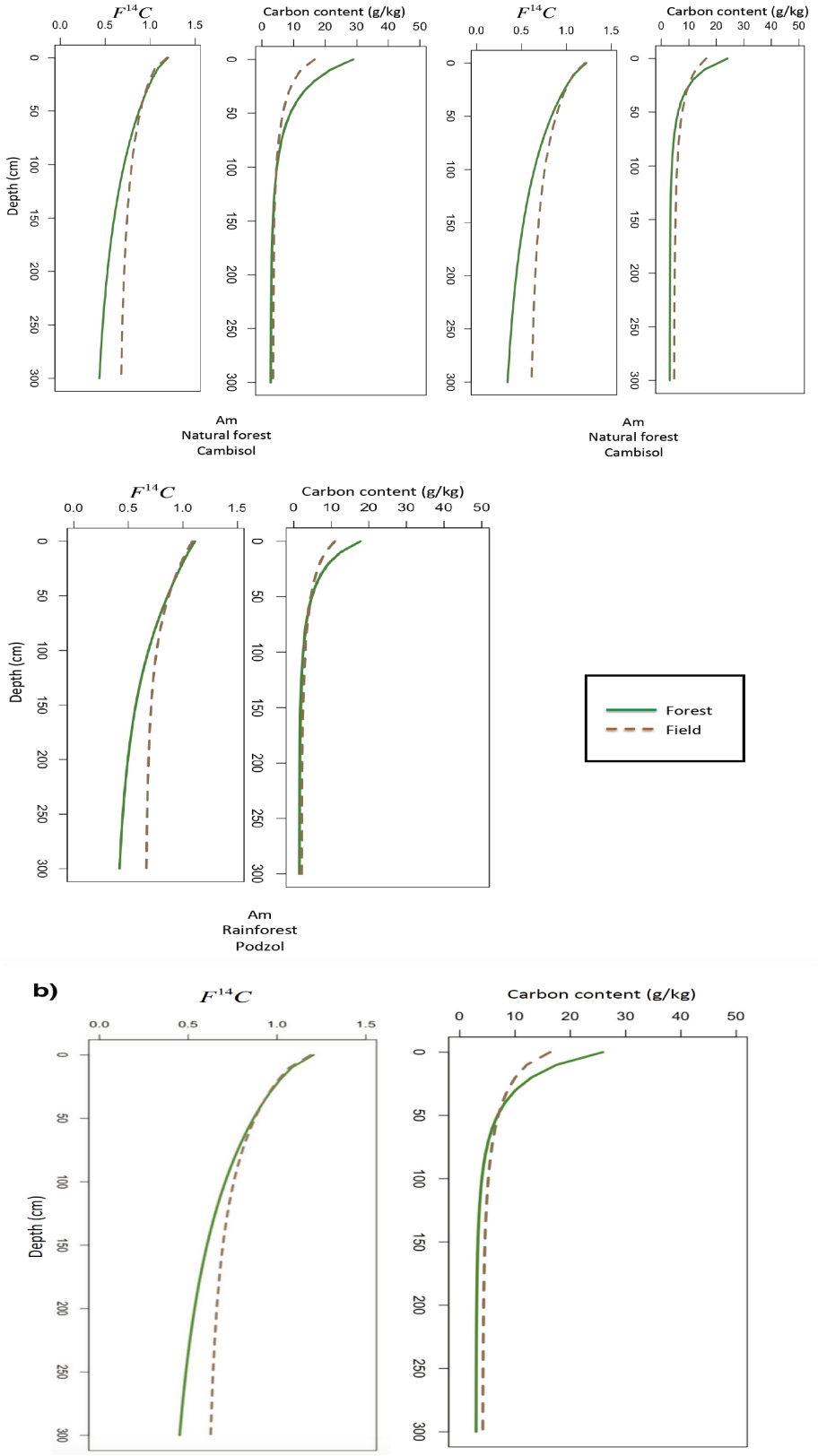


Figure 5.32: Soil radiocarbon and carbon content profiles of each of the nine Brazilian natural forest profiles under rather equatorial climate before (continuous green line) and after the conversion into cropland (dashed brown line) (panel a). The average profile characterized by 25.6 °C of Mean Annual Temperature, 2057 mm of Mean Annual Precipitation and 1.56 of Aridity Index is given in panel b.

A slight decrease from 1.20 to 1.18 is observed for the topsoil radiocarbon when converting the forest to cropland associated to a clear decrease in the topsoil carbon content from 26 g/kg for the tropical forest to 16 g/kg when converting forest to croplands (Figure 5.32). Tropical forest brings more fresh carbon into the topsoil (2200 tons per km^2 per year, Table 5.18) than the cultivated land (650 tons per km^2 per year) as it is characterized by a higher Net Primary Productivity (NPP) (Table 5.18) and shallow rhizosphere.
A significant increase in the deep $F^{14}C$, from 0.45 to 0.58 is observed at deep layers. In other words, the residence time of the carbon at deep layers is greater for tropical rainforest than for cultivated land. But no real change occurs for carbon content in depth.
These results are in line with Balesdent et al. (2018) study that shows that land use for crops reduces the incorporation of carbon into the soil surface layer but not into deeper layers. Our results add further elements to the discussion by Balesdent et al. (2018) by showing that organic matter in our current soils is the legacy of its management by several generations of farmers. The much higher $F^{14}C$ in depth will impact the global carbon cycle over a long period of time with a more rapid return of stored (and thus not stocked) carbon to the atmosphere.

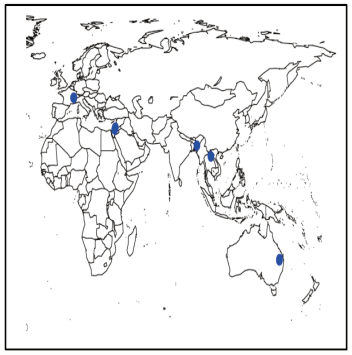
Land type	Average NPP (tons per km^2 per year)
Algal beds and reefs	2500
Tropical rainforest	2200
Swamp and marsh	2000
Estuaries	1500
Temperate deciduous forest	1200
Boreal forest	800
Cultivated land	650
Temperate grassland	600
Continental shelf	360
Tundra and alpine	140
Desert and semi-desert	90

Table 5.18: The variation of the Net Primary Productivity (NPP) per unit area according to the land type (Jackson et al., 1997).

5.5.1.2 Reforestation of temperate cropland and pasture leads to a higher carbon stock on short and long-term duration

Afforestation is commonly regarded as a mitigation solution to address climate warming thanks to an a priori high sequestration potential. However many studies have reported contradictory findings: afforestation results either in a decrease (Garcia-Franco et al., 2014; Li et al., 2014; Wiesmeier et al., 2009), an increase (Nave et al., 2013) or no clear effect (Eclesia et al., 2012) in soil organic carbon stocks. Many factors have been highlighted to explain the extent of change in the soil carbon stock: site preparation, site management, fire, time elapsed since the conversion, previous land use, climate, type of soil, etc. but no univocal relation appears to link these factors to organic carbon content. Recently, Song et al. (2018) reported the increase in land surface covered by forest in temperate regions due to loss of agricultural land and loss of pasture. What will be the consequence of this return of forest in temperate regions?

To address this issue, we selected the ten sites available in the database that met both temperate climate and ecosystem type equal to field (cropland - three profiles) or cultivated grassland (pasture - seven profiles). We are positioned here in the case of reforestation following an agricultural decline rather than the afforestation of a region of weak vegetation (mostly savannah).



We followed the same procedure as previously described: estimating a mean profile of the ten profiles in their initial condition (field or cultivated grassland), then simulating the conversion towards natural forest for all ten profiles and estimating a mean profile of the ten converted profiles.

Soil type	Vertisol	Vertisol	Cambisol	Vertisol
Land use	cultivated grassland	cultivated grassland	field	field
MAT/Dif_T(°C)	19.7/14.2	19.5/14.2	22.5/7.9	19.6/13.8
MAP(mm)	681.6	681.6	1641.4	517.5
Aridity Index	0.40	0.40	1.35	0.36
min_P (mm)	28.3	28.3	13.8	0
Sampling year	1997	1997	2003	1986
Reference data	Krull et al., 2003	Krull et al., 2003	Rumpel et al., 2008	Becker-Heidmann et al., 2002
Country	Australia	Australia	Laos	Israel

Soil type	Vertisol	Luvisol	Fluvisol
Land use	cultivated grassland	cultivated grassland	field
MAT/Dif_T(°C)	18.8/15.2	6.8/17.4	24.1/10.2
MAP(mm)	694.5	1120.9	3047.9
Aridity Index	0.5	1.59	2.60
min_P (mm)	0	82.5	9.1
Sampling year	1985	1980	2009
Reference data	Becker-Heidmann et al., 2002	Balesdent et al., 1982	Laskar et al., 2012
Country	Israel	France	India

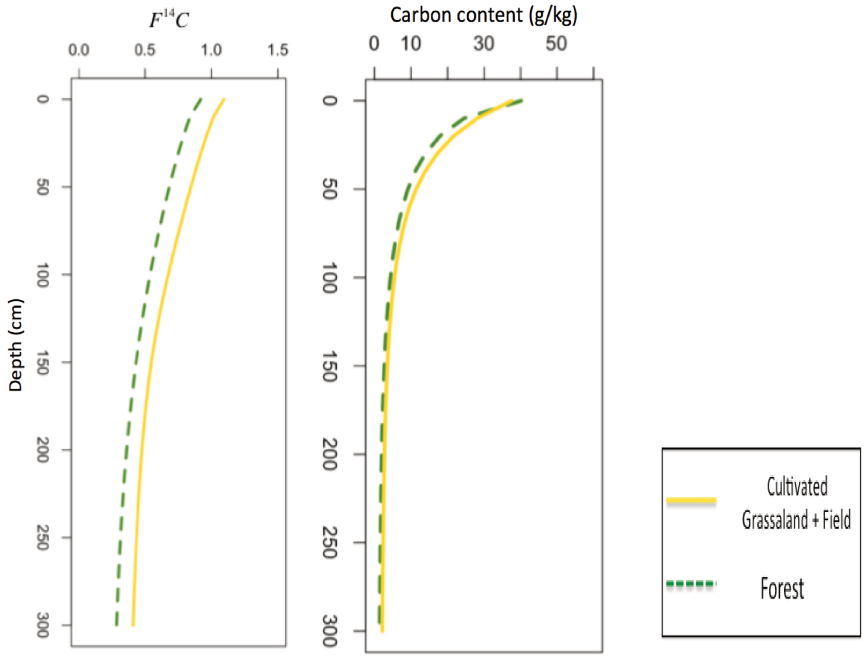


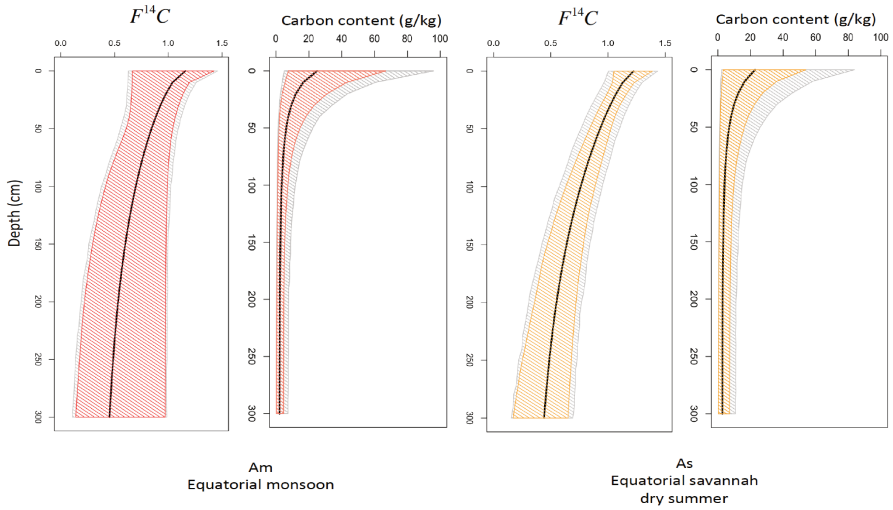
Figure 5.33: The soil radiocarbon and soil carbon content dynamics after (green dashed curve) and before (yellow curve) the conversion of field and cultivated grassland to natural forest for the average temperate profile.

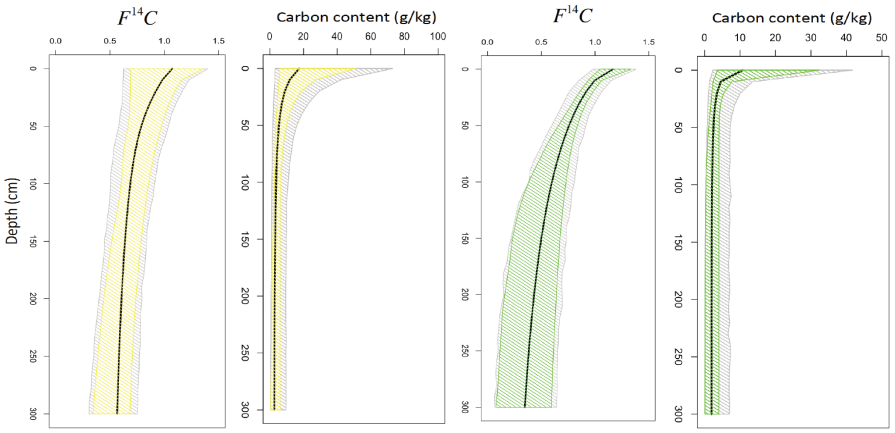
The agricultural decline observed in temperate regions leads to forest regeneration that impacts soil carbon dynamics. Topsoil carbon content slightly increases for the upper topsoil from 37 to 40, decreasing a bit thereafter with depth. Change is more noticeable on the $F^{14}C$ profile. A constant shift towards lower $F^{14}C$ is observed whatever the depth when the aboveground vegetation changes from crop or pasture to forest. It changes from 1.09 to 0.92 for topsoil and from 0.41 to 0.28 for deep soil. This lower $F^{14}C$ reflects an increase in soil carbon turnover and thus a better preservation of soil organic carbon. This is a good result from the global carbon point of view. However, in absolute terms, agricultural decline is not a good point for food safety and autonomy and sometimes not good either for the short food supply chain. The evaluation must be done by integrating the system as a whole.

5.5.2 Impact of climate change

In order to project the impact of the current global warming on soil carbon dynamics, we first plotted the average profile for each of the Köppen climate classifications indicated in Table 5.34. To do so, we decided to use the Full Bayesian model of soil radiocarbon dynamics (inclusion of all climate and environmental predictors) rather than the best sub model as climatic variables were not always selected for all of the latent variables of both $F^{14}C$ and carbon content profiles.

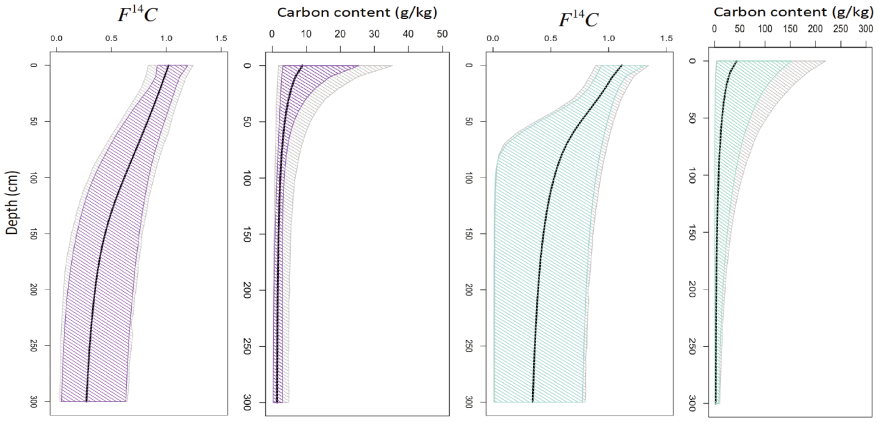
Here also we used the same procedure as previously. A posterior mean estimation profile was obtained from the n profiles from the database that correspond to the Köppen-Geiger climate sub-group (see Table 5.34 to get n). Most of the Köppen-Geiger sub-groups are then presented. Three sub-groups are missing equatorial rainforest climate (Af), desert climate (BW) and frost climate (EF). Simulation was obtained for both $F^{14}C$ and carbon content. The resulting mean estimation profiles are shown in Figure 5.34. A comparison between the average profiles according to the different Köppen-Geiger climate sub-group is given in Figure 5.35.





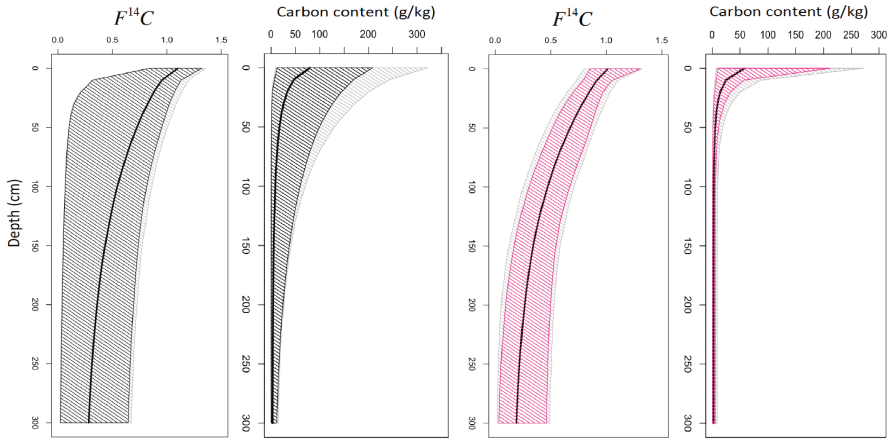
Aw
Equatorial savannah
dry winter

Bs
Steppe climate



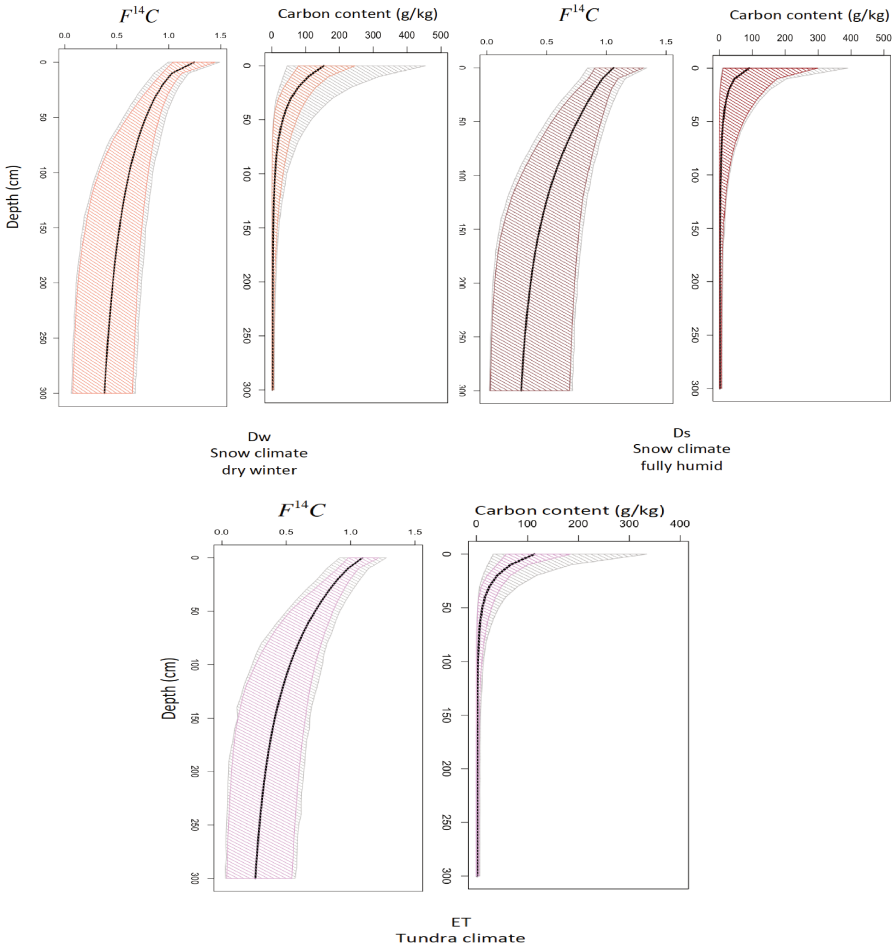
Cs
Warm temperate
Dry summer

Cw
Warm temperate
dry winter



Cf
Warm temperate
fully humid

Ds
Snow climate
dry summer



Type	Description	Criterion	profiles nb
A	**Equatorial climates**	**$T_{min} \geq 18°C$**	**20**
Am	Equatorial monsoon	MAP $\geq$ 25 (100 – Pmin)	10
As	Equatorial savannah with dry summer	$P_{min} \leq$ 60mm in summer	4
Aw	Equatorial savannah with dry winter	$P_{min} \leq$ 60mm in winter	6
B	**Arid climates**	**Pann $\leq 10\ P_{th}$**	**1**
BS	Steppe climate	$P_{ann} \geq 5\ P_{th}$	1
C	**Warm temperate climates**	**-3 °C $\leq$ Tmin $\leq$ +18°C**	**70**
Cs	Warm temperate climate with dry summer	$P_{Smin} < P_{Wmin}$ $P_{Wmax} \geq 3\ P_{Smin}$ $P_{Smin} \leq 40$ mm	5
Cw	Warm temperate climate with dry winter	$P_{Smin} \geq P_{Wmin}$ $P_{Smax} \geq 10\ P_{Wmin}$	18
Cf	Warm temperate climate, fully humid	neither Cs nor Cw	47
D	**Snow climates**	**$T_{min} \leq$-3 °C**	**38**
Ds	Snow climate with dry summer	$P_{Smin} \leq P_{Wmin}$ $P_{Wmax} \geq 3\ P_{Smin}$ $P_{Smin} \leq 40$ mm	15
Dw	Snow climate with dry winter	$P_{Wmin} < P_{Smin}$ $P_{Smax} > 3\ P_{Wmin}$	4
Df	Snow climate, fully humid	neither Ds or Dw	19
E	**Polar climates**	**$T_{max} \leq 10°C$**	**2**
ET	Tundra climate	$0°C \leq T_{max} \leq 10°C$	2

Figure 5.34: Influence of climatic variables on estimated average $F^{14}C$ illustrated for Köppen-Geiger climate classes summarized in the Table below the panels. The dark line represents the average profile relative to the Köppen-Geiger climate class. The colored band corresponds to the between-site variability of soil profiles with similar Köppen-Geiger climate classes. The gray band corresponds to the confidence within the same site.

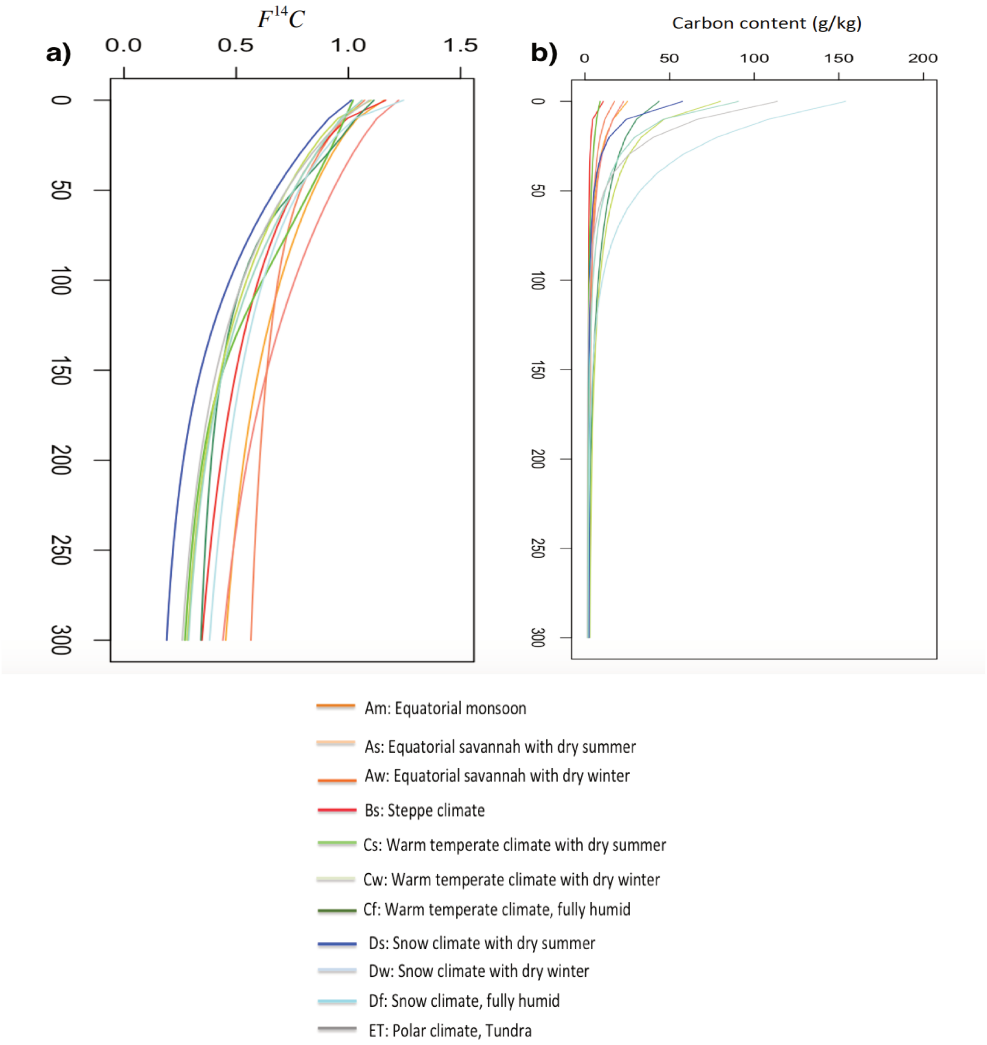
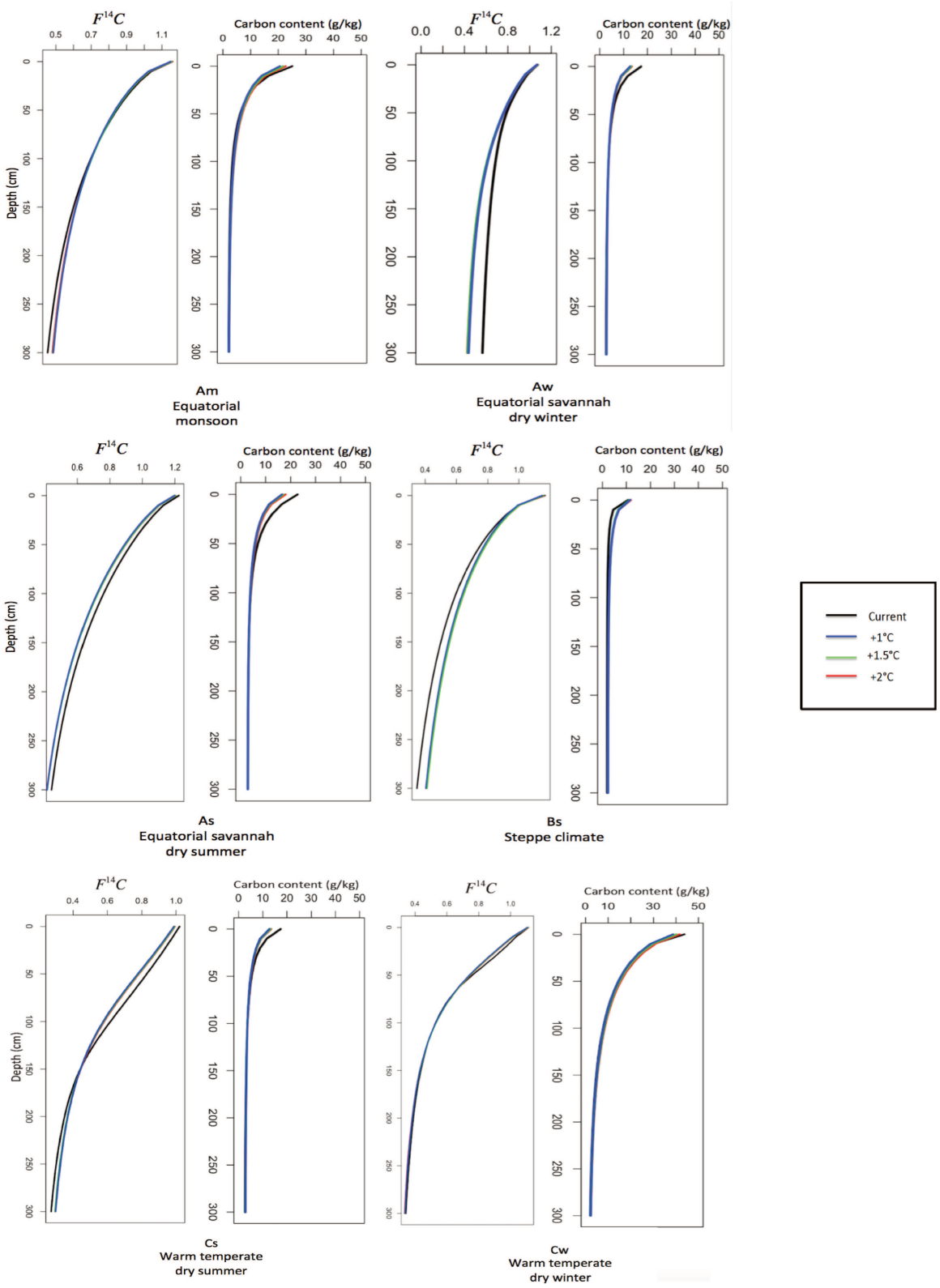


Figure 5.35: Superposition of the average profiles obtained for all the Köppen-Geiger climate sub-groups from the database for $F^{14}C$ (panel a) and carbon content (b).

The succession of $F^{14}C$ profiles as shown in Figure 5.35 highlights global feature differences according to the climate. The colors used for Figure 5.35 highlight some general features. Snow climate (blue derived colors in Figure 5.35) roughly shows a lower $F^{14}C$ than Equatorial climate (red derived colors), leaving the Temperate climate in between. This means that residence time is higher, organic matter is better preserved in cold climates than in equatorial ones. This is clearer for deep soil than for topsoil for which differences remain tiny. Deep carbon is more stable and resistant under a snow climate with dry summer (dark blue curve) and shows the lower mean residence time under equatorial savannah with dry winter climate (coral curve). The same trend is recorded for topsoil.

To address the global warming issue, an increment of 1°C, 1.5 °C and 2°C was applied to the 11 mean profiles that correspond to the 11 climate subgroups from the database. The same procedure as previously described was applied: projecting each individual profile in a climate with MAT +1°C, and evaluating the mean profile climate sub-group by climate sub-group. All the results are shown in Figure 5.36.


$F^{14}C$
Carbon content (g/kg)
Depth (cm)
Am
Equatorial
monsoon
Aw
Equatorial savannah
dry winter
As
Equatorial savannah
dry summer
Bs
Steppe climate
Cs
Warm temperate
dry summer
Cw
Warm temperate
dry winter
Current
+1°C
+1.5°C
+2°C

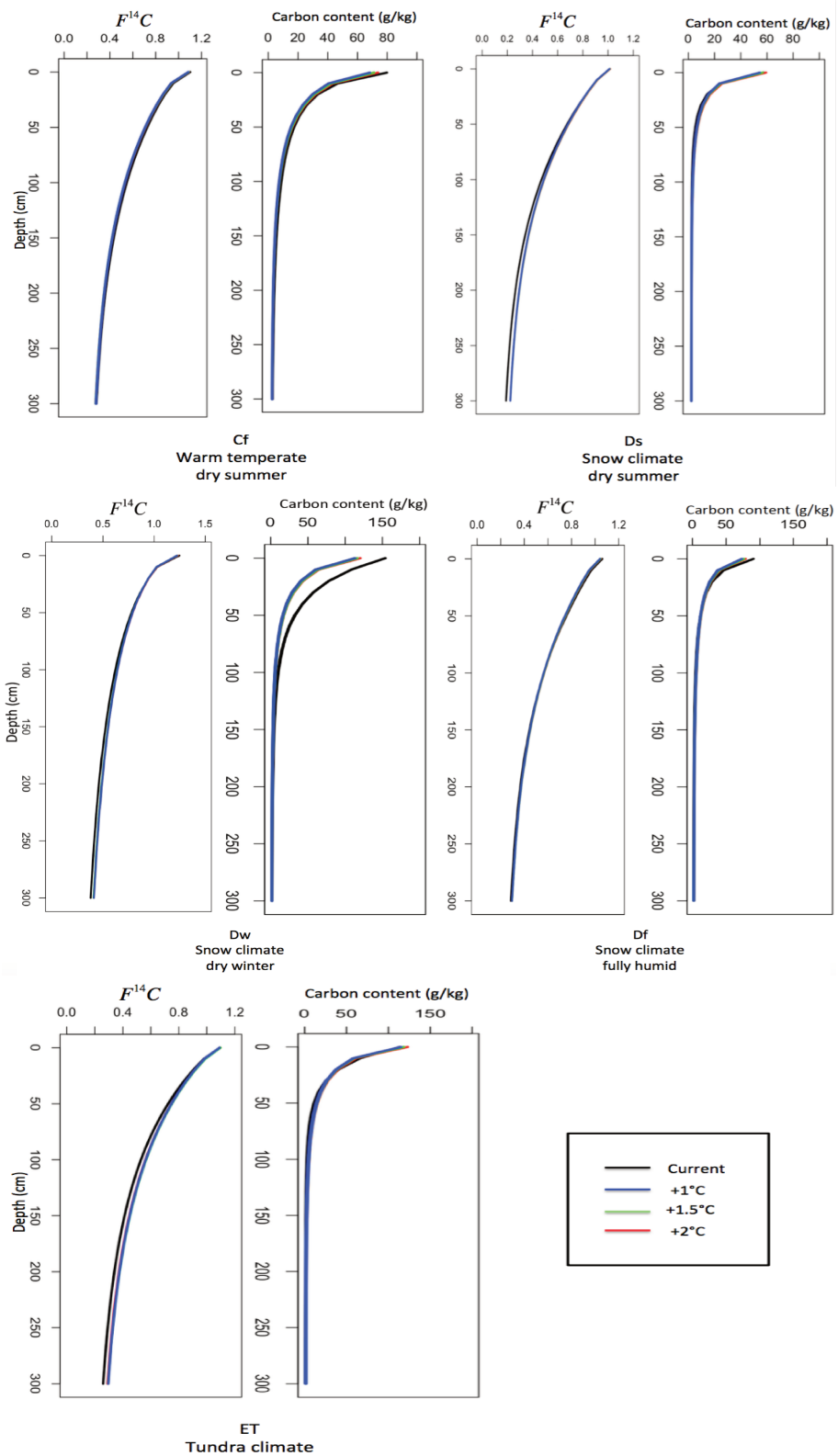


Figure 5.36: Average profiles of $F^{14}C$ and soil carbon content under the current Mean Annual Temperature (MAT, black curve) and predicted for an increment of 1°C (red curve), 1.5°C (green curve) and 2°C (blue curve) of MAT. Each panel is relative to a sub-group of the Köppen-Geiger climate classification

In Figure 5.36, for some profiles, the diffrence between +1°C, +1.5°C and +2°C is minor. In fact, this result can be due to the prediction of the profiles with +1°C, +1.5°C and +2°C from the marginal distribution without taken into account of the observed mesures. Thus, the uncertainty is greater and the signal corresponding to the Mean

Annual Temperature can be lost.

The impact of global warming as shown in Figure 5.35 differs in extent for both $F^{14}C$ and carbon content profiles depending on the climate subgroup. According to the Bayesian selection, an increase in MAT results in i- a decrease in topsoil carbon content that might be related to an increase in microbial activity and higher soil carbon mineralization, ii- no impact on deep soil carbon content, iii- no real change in top soil$F^{14}C$, iv- variable (from negative to positive, including null) effect on deep soil $F^{14}C$. Some of these mathematical results are well reflected in the distribution shown in Figure5.35. We do indeed find a lower carbon content under equatorial than under snow climate, and no real change in depth. However, the effect on topsoil$F^{14}C$ is noticeable and not as small as the Bayesian selection returned and the effect on deep soil $F^{14}C$, while not univocal, is less variable, and a trend towards a decrease in $F^{14}C$ with increasing MAT is observed.

To address the global warming issue, an increment of 1 °C, 1.5 °C and 2 °C was applied to the 11 mean profiles that correspond to the 11 climate subgroups present in the database. The same procedure as previously described was applied: projecting each individual profile in a climate with MAT +1 °C, and evaluating the mean profile climate sub-group by climate sub-group. All the results are shown in Figure 23 and Table 4 provides key numbers. The following observations can be made:

- It is important to highlight that if an impact is recorded for a change of +1 °C, the impact does not propagate in extent for +1.5 °C and +2 °C. The +1.5 °C and +2 °C profiles are superimposed on the +1 °C profile.
- The impact of warming differs greatly from one sub-group to another. The profiles that evolve the most with the global warming are under Equatorial savannah climate with dry winter, Steppe climate, Warm temperate dry climate with dry summer, Snow climate with dry winter or summer. This means that about half of the climate sub group does not show a noticeable change in carbon profiles associated to global warming.
- Within the impacted profiles, the profiles under a Steppe climate are those that change the most. This sub group is however represented by only one profile from the database. This soil shows a large change in deep soil $F^{14}C$ with an increase of 55% in the current value, from 0.11 to 0.17 resulting in a massive destocking of deep old carbon associated to warming. This result is accompanied by an 17% increase in deep carbon content and a shift towards deeper depth of the incorporation depth for carbon content. No conclusions can be drawn however from a single profile.
- Profiles associated to Equatorial savannah climate with dry summer also evolve with the MAT increase. It results in a 22% decrease in the topsoil carbon content and a shift towards a lower $F^{14}C$ of deep soil (change of 29%, from 0.48 to 0.34). This decrease in $F^{14}C$ is associated to a slight increase in the carbon content in depth. So a MAT increase seems to imply a higher plant input in topsoil, likely associated to higher vegetation NPP and surprisingly to a higher residence time in depth. Does this reflect a shift towards a profile with less bomb-peak derived carbon?
- The carbon content profile under snow climate with dry summer shows an increase in deep soil $F^{14}C$, leading to a lower residence time in depth. This is associated to a small decrease in carbon content. No change is recorded in topsoil.
- The carbon content profile under Warm temperate climate with dry summer shows odd changes with a huge increase in the incorporation change (from 8.8 to 15.4cm). In parallel, an increase in deep soil $F^{14}C$ is recorded, *i.e.*, a 17% lower residence time.

In summary, the impact of an increase in MAT is visible on soil carbon profiles and noticeable for half of the climate sub-groups. However, there is no univocal signal associated to a MAT increase in our data analysis. Impact can result either in an increase, a decrease or no change for both the $F^{14}C$ profile and carbon content profile and

for all of the latent variables. There is no global signal and special attention should be paid at the regional scale of equatorial climate with a dry season that shows a better storage capacity.

Subclimate group	Equatorial monsoon				Equatorial savannah dry summer				Equatorial savannah dry winter				Steppe climate			
short name	Am				As				Aw				Bs			
MAT	Current	+1°C	+1.5°C	+2°C	Current	+1°C	+1.5°C	+2°C	Current	+1°C	+1.5°C	+2°C	Current	+1°C	+1.5°C	+2°C
$\phi 1$	0.29	0.332	0.335	0.336	0.25	0.234	0.235	0.245	0.48	0.340	0.342	0.348	0.11	0.171	0.175	0.180
ϕ_2	1.16	1.156	1.152	1.149	1.224	1.204	1.202	1.197	1.072	1.076	1.073	1.070	1.16	1.157	1.156	1.151
ϕ_3	151.32	151.30	151.59	151.16	168.6	164.1	163.76	163.17	103.6	109.4	109.8	108.9	172	180.18	179.58	180.68
ω_1	1.99	2.05	2.06	2.05	2.76	2.84	2.86	2.86	2.53	2.76	2.76	2.77	1.96	2.29	2.31	2.29
ω_2	25.06	22.87	21.91	20.77	22.77	18.02	17.01	16.45	17.3	13.5	12.9	12.5	10.7	11.5	11.7	10.8
ω_3	39.97	45.40	42.21	45.22	39.04	34.87	34.78	34.88	53.1	45.8	46.4	46.1	20.16	26.28	25.89	26.47

Subclimate group	Warm temperate dry summer				Warm temperate dry winter				Warm temperate fully humid				Snow climate dry summer			
short name	Cs				Cw				Cf				Ds			
MAT	Current	+1°C	+1.5°C	+2°C	Current	+1°C	+1.5°C	+2°C	Current	+1°C	+1.5°C	+2°C	Current	+1°C	+1.5°C	+2°C
$\phi 1$	0.184	0.216	0.216	0.216	0.267	0.260	0.263	0.268	0.185	0.182	0.185	0.187	0.127	0.158	0.159	0.163
ϕ_2	1.021	0.997	0.993	0.988	1.113	1.113	1.110	1.106	1.101	1.091	1.088	1.084	1.014	1.019	1.015	1.012
ϕ_3	147	137.08	137.66	136.47	111.05	107.87	107.25	106.80	119.91	117.06	116.67	116.37	106.94	109.56	108.78	108.40
ω_1	1.29	1.29	1.29	1.30	1.55	1.54	1.55	1.55	2.33	2.25	2.25	2.25	2.34	2.25	2.25	2.25
ω_2	8.84	15.41	14.82	13.39	43.67	41.61	40.27	38.72	79.83	73.97	71.12	68.70	57.51	59.25	56.87	54.19
ω_3	62.13	54.15	54.17	54.41	36.55	39.78	39.96	40.08	31.35	31.06	31.12	31.15	20.56	20.11	20.18	20.11

Subclimate group	Snow climate dry winter				Snow climate fully humid				Polar climate Tundra			
short name	Dw				Df				ET			
MAT	Current	+1°C	+1.5°C	+2°C	Current	+1°C	+1.5°C	+2°C	Current	+1°C	+1.5°C	+2°C
$\phi 1$	0.145	0.182	0.182	0.184	0.188	0.197	0.199	0.202	0.151	0.179	0.184	0.183
ϕ_2	1.245	1.233	1.228	1.224	1.06	1.05	1.045	1.041	1.093	1.099	1.094	1.091
ϕ_3	154.17	156.98	156.66	155.86	129.63	129.42	129.26	128.42	120.23	122.48	123.18	122.75
ω_1	1.68	1.65	1.65	1.66	1.79	1.83	1.82	1.82	1.31	1.24	1.25	1.27
ω_2	153.82	119.84	116.80	113.01	90.53	78.98	76.19	73.26	113.59	123.15	117.81	115.02
ω_3	33.70	24.93	24.63	24.85	23.14	23.23	23.38	23.20	21.91	21.22	21.56	21.10

Table 5.19: The latent values for both $F^{14}C$ and soil carbon profile corresponding to current temperature and an increase of the MAT by +1°C, +1..5°C and +2.5°C. ϕ_1 and ω_1 refer to deep $F^{14}C$ and carbon content respectively, ϕ_2 and ω_2 represent the topsoil $F^{14}C$ and soil carbon content and finally ϕ_3 and ω_3 underline the $F^{14}C$ and carbon content incorporation depth.

CHAPTER 6

CONCLUSIONS & PERSPECTIVES

6.1 Summary and Conclusions

Soil carbon is a solution to mitigate global warming

Understanding the dynamics of soil carbon is a major challenge, especially as the IPCC pointed out the large uncertainty on the soil carbon stock and its potential impact on future climate change. **The large capacity for carbon exchanges with the atmosphere, the huge uncertainties about the response in soil carbon to global changes in climate and land use practices and lastly the fact that soil carbon is the only pool that humans can manage, show the crucial global interest of better understanding the fate of soil carbon**. Increasing the soil carbon stock is foreseen as a solution to mitigate global warming but this is relevant only if the storage is sustainable.

Incomplete physical representation of soil carbon dynamics by mechanistic models

Several mechanistic models have been developed to better understand the behavior of soil carbon based on expressing the physical processes by differential equations. Among these mechanistic models, one can mention the Roth-C Model, and the CENTURY model. These soil models are also integrated into Land Surface Models, such as ORCHIDEE, which is the land component of the Institut Pierre Simon Laplace (IPSL) Earth System Model. However, **the incomplete view of the physical protection processes (spatial inaccessibility, encapsulation in organic macromolecules, etc.), the disregard of deep soil carbon layers and the parametrization of mechanistic models under specific management and climatic conditions highlight that there are still missing unknown terms in the soil carbon dynamics representation**. Furthermore, improving the mechanistic models by integrating the missing physical concepts requires years and years of research work.

Faced with all these challenges, how can statistical meta-analysis help soil scientists?

Statistical meta-analysis helps decision making

Statistical modeling can provide faster responses to the scientific questions of today and the near future about soil carbon and it can consider uncertainties.

Using Statistical modeling to explore a soil carbon database

Before speaking about the statistical model, the database used for this study required long-term collecting and updating by Christine Hatté from 85 articles in the soil science, archaeology and paleoclimatology fields. The database contains 343 soil carbon profiles distributed roughly over the entire globe. Each profile displays radiocarbon and carbon content measurements at different depths. Furthermore, climatic and environmental conditions were reported (temperature, precipitation, soil type, ecosystem type, etc.). Unfortunately, the experimental design usually employed in a meta-analysis is not balanced. In our study, the only factors are the soil and ecosystem types. In order to improve the experimental design, we merged some categories that share the same features within the same factor. Whatever the efforts that can be made, we can improve but not avoid this issue since the soil type and ecosystem are naturally associated. The database provides a homogeneous representativeness of intermediate climates, *i.e.* tropical, warm temperate and snow climates, leaving out extreme climates (arid and polar).

The database has the advantage of providing soil radiocarbon and carbon content for each profile. A parallel comparison of these two profiles afforded a twofold vision of soil carbon content and its mean residence time since the radiocarbon is characterized by its radioactive decay. In addition, among the profiles of radiocarbon (131

profiles) and carbon content (125 profiles) remaining after database cleaning (removing paleosol, profiles with unknown information, etc.), 58 profiles have measurements for depth levels exceeding 100 cm. Thus, the deep soil carbon is well represented in the database.

Bayesian inference better represents model uncertainties

The variation in both soil radiocarbon and soil carbon content with depth are separately represented by a hierarchical non linear model with latent variables which in turn are linked to the climatic and environmental potential predictors. The estimation of unknown model parameters was done using Bayesian inference. Bayesian inference gives us the possibility to integrate expert advice about the model parameters. Unfortunately, sometimes, it is not easy to provide an informative prior, either because no prior beliefs on model parameters are available or because the soil scientist prefers to be objective and see what we can learn from data. Thus, for each soil carbon profile, Bayesian inference makes it possible to compute the credible intervals which highlight uncertainties within and between soil carbon profiles. The statistical model has the advantage of predicting the soil carbon profile for both radiocarbon and carbon content for sites where no data have been collected. We only need to know the climatic and environmental site conditions.

Bayesian selection methods provide a probabilistic judgment about the contribution of climatic and environmental factors to soil carbon dynamics

As a matter of fact, soil scientists have an idea about the potential climatic and environmental factors that can impact the soil carbon dynamic but they do not know, in the first place how to prioritize these factors by their effects on soil carbon dynamics and in the second place there is still debate on some issues. For example, the soil science community is divided on the question of the soil carbon sensitivity to temperature changes. Some soil scientists think that the topsoil carbon is more sensitive to temperature changes than the deep soil carbon, others consider that the top and deep soil have the same sensitivity to temperature changes and the remaining ones think that the deep carbon is more influenced by temperature changes than the topsoil.

The first goal achieved in my thesis was to provide a probabilistic judgment on inclusion of the climatic and environmental factors which have a physical interpretation in the latent layer models. I first explored the Stochastic Search Variable Selection (SSVS). This approach introduced by George and McCulloch (1993), was designed to select numerical predictors in the framework of a linear model. However, even after adapting the SSVS to support the inclusion of the categorical predictors such as the soil and ecosystem types used in our study, the results obtained on artificial data, generated according to the proposed statistical model for soil carbon dynamics, proved that the detection of some significant categorical predictors with SSVS can be misleading. Thus, **the second challenge achieved was to investigate other Bayesian selection approaches appropriate for categorical predictors**. To ensure the best selection of the categorical and numerical covariates, three Bayesian Selection approaches were explored: Bayesian Group Lasso with Spike and Slab priors (BGL-SS), Bayesian Sparse Group Selection (BSGS) and Bayesian Effect Fusion model-based clustering (BEF). In addition to selecting categorical predictors, the BSGS also provides a selection by level within the same predictor and the BEF makes it possible to merge the levels within the same predictor having the same effect on the variable on interest.

The best sub-sets of climate and environmental factors selected were obtained by comparing the previous selection methods based on the Bayesian selection Criteria such as the Deviance Information Criterion (DIC), the 5-fold cross validation, etc.

The best sub-sets of selected climate and environmental factors show very interesting findings:

1. The non detection of the $F^{14}C$ of the sampling year for topsoil $F^{14}C$ is explained by a non representative distribution of database profiles by sampling year.

2. The deep soil layer radiocarbon is identified as more sensitive to temperature changes than the topsoil radiocarbon.

3. The Mean Annual Precipitation affects both topsoil and incorporation depth in radiocarbon and carbon content profiles.

4. The Aridity Index is only an influential predictor for topsoil and for incorporation depth of the carbon content profile.

5. The result of merging soil types for both topsoil $F^{14}C$ and deep $F^{14}C$ underlines that the $F^{14}C$ is mainly dominated by the climate/vegetation and soil texture at the topsoil and by clay content for deep layers.

6. Soil type and land use affect both topsoil and deep layers of radiocarbon and carbon content profiles.

Besides these very encouraging results, it appears that the impact of a climatic or categorical factor on the latent variable representing separately the incorporation depth for $F^{14}C$ and soil carbon content (ϕ_3 and ω_3 latent variables) is highly overestimated.

The statistical model provides prediction in a context of global changes

The third achieved goal was to illustrate how the statistical model can be used to predict the changes occurring in the soil carbon profile when the land use or climatic conditions change. Here we show the impact of land use change with two examples: 1- when equatorial forest is replaced by cropland and 2- when cultivated grassland and field are replaced by cultivated forest in a temperate region.
The first example shows that the conversion of equatorial forest to cropland impacts both topsoil soil carbon content and deep radiocarbon dynamics. After conversion, the topsoil carbon content decreases from 26 g/kg to 16 g/kg but shows a decrease in the mean residence time of deep soil carbon (deep soil radiocarbon increases from 0.45 to 0.58). For the second example, the reforestation of temperate cropland and pasture yields a higher carbon stock and long-term duration. A third example highlights the impact of global warming on soil carbon dynamics. Radiocarbon and carbon content profiles plotted according to the Köppen-Geiger climate classification showed that the deep carbon is more stable under a snow climate with dry summer and has a lower mean residence time under an equatorial savannah with dry winter climate. The temperature increment is more visible on the carbon profile on going from the current temperature to an increase of +1 °C than on going from an increase of +1.5 °C to +2 °C. The impact of temperature increase may be negligible for soil in a temperate climate but considerable for both deep and topsoil in a snow climate.

Finally, there are still some ideas and objectives that I have not had time to look at in detail and that I will discuss in the ”Perspectives” section.

6.2 Improvements and Perspectives

Several propositions and possibilities can be investigated in order to improve the structure of the statistical model and to a further outlook for the use of the statistical model. I distinguish here, the following points:

- **A bivariate Gaussian Process for both radiocarbon and carbon content dynamics can better express model uncertainties**

Observation of all soil $F^{14}C$ and carbon content profiles from the database (Figure 6.1) underlines a dependence between the measurements at different depths.

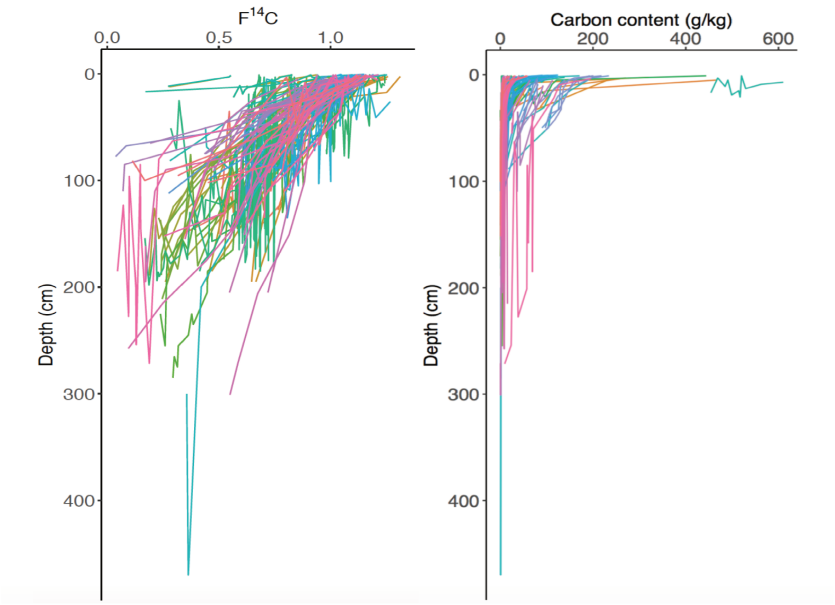


Figure 6.1: All $F^{14}C$ and soil carbon content database profiles.

In fact, the radiocarbon and soil carbon content measurement observed at a depth of 10 cm are expected to be higher than a measurement observed at a depth of 20 cm. In the current statistical models, the measurements of the quantities of interest, radiocarbon and soil carbon content, are considered as independent. In order to better represent model uncertainties, a Gaussian model under the same non linear hierarchical structure can be useful to take into account the dependency between observations. Thus, if we suppose a site s with $F^{14}C_s$ = $(F^{14}C_s(z_1),\ldots,F^{14}C_s(z_{m_s}))$ measurements evaluated at m_s depth levels $z \in \mathbf{R}^{m_s}$, the Gaussian Process (GP) considers that the marginal distribution over that finite subset has a multivariate distribution and that it is characterized by a mean m and covariance function k (that need to be defined).

$$F^{14}C_s \sim GP(m_{F^{14}C}, k)$$

Here we can imagine a mean function $m_{F^{14}C}$ that has the same non linear structure with random effects as the statistical model proposed before:

$$m_{F^{14}C} = \phi_1(s) + (\phi_2(s) - \phi_3(s)) * exp(-(z/\phi_3(s))^{\phi_4(s)}$$

Several covariance functions k have been proposed in the literature (Rasmussen, 2004). A very popular function is the exponential quadratic kernel that has two components to calibrate: the marginal standard deviation of the stochastic process α and the length scale l, roughly the distance one must move in input space before the function value can change significantly.

$$k(z_i, z_j) = \alpha^2 exp\left(-\frac{1}{2l^2}(z_i - z_j)\right)$$

In addition to their flexibility, GP also easily provide prediction, even in the absence of strong physical models. In fact, once the mean "m" and the covariance function "k" are defined, the posterior prediction distribution of

the soil $F^{14}C$ corresponding to new depth levels is ensured thanks to the following joint distribution:

$$\begin{bmatrix} F^{14}C \\ F^{14}C^* \end{bmatrix} \sim \left(\begin{bmatrix} m_{F^{14}C} \\ m_{F^{14}C^*} \end{bmatrix}, \begin{bmatrix} k & k^* \\ k^{*T} & k^{**} \end{bmatrix} \right)$$

we can define the conditional distribution $F^{14}C|F^{14}C^*$ as a predictive distribution for the new point $F^{14}C^*$ as:

$$F^{14}C^*|F^{14}C \sim G(m_{F^{14}C^*} + K^{*T}K^{-1}(F^{14}C - m_{F^{14}C}), k^{**} - k^{*T}k^{-1}k^*)$$

The Bayesian formulation has the advantage of reducing the computational time by proposing a prior distribution on the variance scale α and the length scale l. I tried to adjust a Gaussian Process model on soil radiocarbon dynamics but further investigation was not done because of lack of time. Figure 6.2 underlines the flexibility of the GP for data interpolation: the mean curve in black tries the best to go through all the measurement points. As expected, the uncertainty increases when no data are observed.

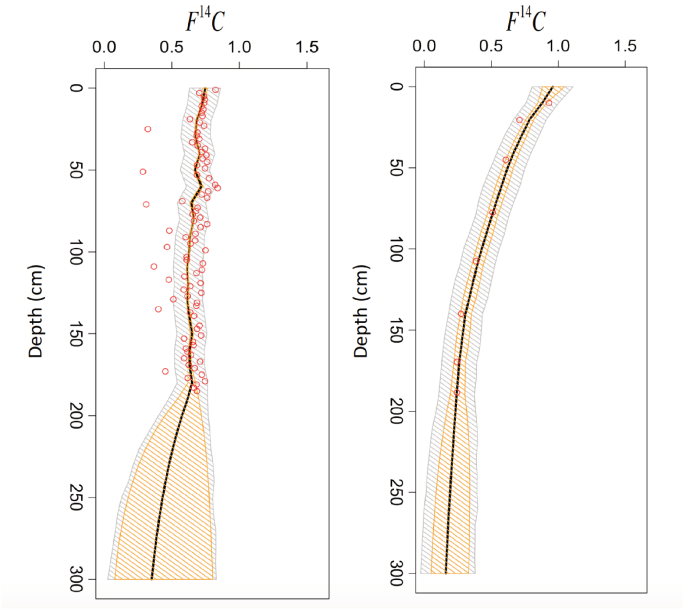


Figure 6.2: Prediction of two $F^{14}C$ profiles chosen from the database according to the Gaussian Process statistical modeling of soil radiocarbon dynamics. The orange band presents the uncertainty inter profile, while the gray band represents the uncertainty intra profile. The red points highlight the real observed radiocarbon measurements.

Here, I have only discussed the application of Gaussian Process to the soil radiocarbon profile, but the same approach can be used for soil carbon content dynamics.

Added to that, why not consider one statistical model with two joint output responses: soil radiocarbon and soil carbon content? For this purpose, one may wonder if there exists a link between soil radiocarbon activity and soil carbon contents. For levels below the upper ones, *i.e.* for which the mean carbon age is older than the bomb peak, soil radiocarbon activity decreases as the soil carbon content decreases. The reverse occurs for the upper levels whose mean age is younger than 50 yrs and thus that are marked by the bomb peak. This case is not well represented in our database, however, so the classical decrease of $F^{14}C$ with depth together with the decrease in organic carbon is the global and most common fate in the database.

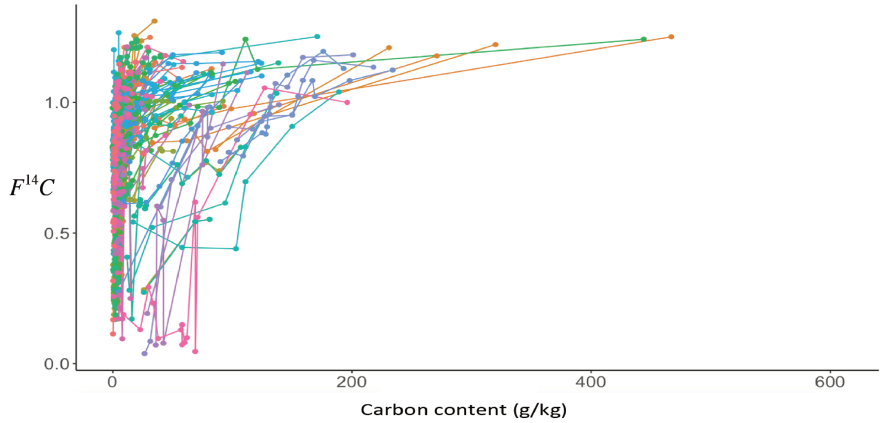


Figure 6.3: The soil radiocarbon activity variation against the variation in soil carbon content based on 125 database profiles.

For statistical modelling, I suggest that the use of a bivariate Gaussian model (Daira et al., 2016) can be helpful to describe the dependency between the radiocarbon and soil carbon content measurements points with depth and the relationship between $F^{14}C$ and carbon content responses.
Thus, for each site $s \in [1:S]$ with $z \in [1, m_s]$ measurements, let us consider the following vector $Y_s = (Y_{1,s}^T, Y_{2,s}^T)^T$ with $Y_{i,s} = (Y_i(z_1), \ldots, Y_i(z_{m_s})$ for i = 1, 2. In other words, $Y_{1,s}$ and $Y_{2,s}$ define the $F^{14}C$ and the soil carbon content measurements respectively. The bivariate model for the site s could be written as follows:

$$\begin{bmatrix} Y_{1,s} \\ \hline Y_{2,s} \end{bmatrix} \sim N\left(\begin{bmatrix} M_{1,s} \\ \hline M_{2,s} \end{bmatrix}, A \otimes R \right)$$

where, $M_{1,s} = (M_1(z_1), \ldots, M_1(z_{m_s}))$ and $M_{2,s} = (M_2(z_1), \ldots, M_2(z_{m_s}))$ vectors are the respective means for radiocarbon and soil carbon content and they are defined as follows:

$$M_1(z_j) = \phi_1(s) + (\phi_2(s) - \phi_1(s)) * exp\left(-\left(\frac{z_j}{\phi_3(s)} \right)^{\phi_4(s)} \right) \quad j = 1, \ldots, m_s$$

$$M_2(z_j) = \eta_1(s) + (\eta_2(s) - \eta_1(s)) * exp\left(-\frac{z_j}{\eta_3(s)} \right) \quad j = 1, \ldots, m_s$$

The covariance matrix is defined by $A \otimes R$ as follows:

$$A = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}, \quad R = [e^{-\theta|z_m - z_n|}] \quad m, n = 1, \ldots, m_s$$

[1] here, $\sigma_1^2, \sigma_2^2 > 0$ are the variance parameters for $F^{14}C$ and carbon content respectively and $\theta > 0$ is the correlation decay parameter. The parameter ρ with $|\rho| < 1$ represents the correlation parameter between soil $F^{14}C$ and soil carbon content.

[1] The Kronecker product of the matrix A et R, denoted by $A \otimes R$ is written as:

$$\begin{bmatrix} a_{11}R & | & a_{12}R \\ \hline a_{21}R & | & a_{22}R \end{bmatrix}$$

- **The database can be better managed**

  - Due to lack of time, the analysis of soil bulk density variability required to convert carbon content into soil carbon stock could not be carried out. This step will have to be done to project the results at the global scale. It might consist of studying measured bulk density within the database or of using pedotransfer functions that differ with the type of soil rather than the universal Alexander equation.
  - At the interpretation steps, some profiles were highlighted outside the general trend. These include, in particular, some profiles in the UK under very wet climate. These profiles likely biased the variable selection and highly likely the extent of impact value. A second set of evaluations can be conducted without these "outlier" profiles to get a statistical model that would better mimic the general trend.

- **It is necessary to improve the experimental design of the database**

Another challenge will be to improve the current database. As we noted in Chapter 2 that describes the database and later on for the physical interpretation of the variable selection and fusion, the database is far from being perfect. For example, we don't have the same number of profiles by soil and ecosystem type. 37% of the database profiles are forest while only 8% are defined as cultivated grassland. Even for soil type, 9% of the database profiles are defined as Andosols and 7% as Regosols/Arenosol/Leptosol. However, only 1% of the total continental land area on earth is occupied by Andosols and 22% by Regosols/Arenosol/Leptosol. Furthermore, as soil type and ecosystem are associated, it might also be of interest to divide Gleysol into two categories: tropical and boreal Gleysols. Some of their features are similar as they have the same name but some others, especially interaction with vegetation, are different.
Increasing the number of profiles by soil and ecosystem (land use + vegetation) category is not the only important point, however; there are several crucial features that also need to be addressed such as: the total continental land area occupation by soil and vegetation cover, the association between: soil, vegetation cover and the climate conditions. In addition, the current database does not contain profiles from the arid and hyper arid climate classes. This is a crucial lack, especially for the use of the model in a prediction mode in these regions that are particularly vulnerable to climatic changes. As stated in the database description, the inhomogeneous distribution of the sampling years in the database hampers a good representation of the $F^{14}C$ profile, in particular for the latent variable that mimics the incorporation depth (ϕ_3).
A further point to improve in the database is thus the distribution of the profiles by sampling year. In the current database, the profiles sampled between 1990 and 2000 represent 53% of the database profiles. As a consequence, the atmospheric radiocarbon related to the bomb peak in the early 1960s was not detected as influential for the topsoil $F^{14}C$ nor for the incorporation depth.Thus, why not consider archives samples, such as the ones of the Rothamsted archives? This collection of samples was established by Lowes and Gilbert in 1843. Several thousands of soils collected in the 1920s-1950s are stored in the repository. About 1200 crop and 200 soil samples are added annually to the archive.
Finally, the question remains as to how to consider permafrost in the soil carbon statistical model. In fact, permafrost accounts for about 20% of the Earth's land surface. This type of soil is particular since its temperature subsurface never rises above 0°C for at least one year.

A deeper analysis of impact of the level width on the profile modeling might also help in defining the sampling pattern by pointing out the best compromise between fieldwork and analytical work and the best representation of the profile specificities. It might thus result in advice on the maximum depth to be reached according to the type of soil, for the level thickness according to the depth (likely a finer resolution for top and mid depth soil

and a lower resolution for deep soil).

- **The extrapolation of the Bayesian statistical model developed for the soil carbon content is useful to have a global (or regional) estimation of the soil carbon stock**

The Bayesian statistical models for radiocarbon and soil carbon content dynamics allow us to predict the soil radiocarbon and carbon content profiles of a new site, knowing the corresponding climatic and environmental information. As no measurements is provided for this site, wider credible intervals will be observed for the unknown model parameters.

In the first place, the predicted profiles, when changing the land use or climatic conditions, are obtained without taking into account the observed measurements (Chapter 5.5.2, Section 5.5). Which means that the site is considered as a new site. Thus, new latent variables are generated for the soil radiocarbon and soil carbon content models. These considerations raise the following question: how can we predict the profile with the changes in climatic or environmental conditions, taking into account the measurements observed for the corresponding site? We can imagine that we have to add to the current and estimated latent variables under the ambient climatic and environmental conditions for site s, the change of effect due to replacing a forest by a cropland or to increasing the temperature by 1°C. In contrast, following this proposition does not guarantee the positivity constraint of model latent variables.
In the second place, one can wonder how we can extrapolate the Bayesian statistical model to have a predicted profile of soil carbon content or radiocarbon at regional or global level. With a more complete database and a more balanced experimental design, other doors open. It then becomes possible to apply the statistical model designed on the whole database to models for each type of soil and each climate zone. This would increase the power of the projection capacity of the study. It would make it possible to better decipher the impact of land-use change according to the soil type, and to better project the impact of present global warming according to the world regions. GIS is thus possible.
Here, we are talking about an extrapolation to 3 dimensions: longitude, latitude and depth. Digital Soil Mapping (DSM) using contextual information spatial models in deep learning is very popular and can be used to generate maps (McBratney et al., 2003). In fact, there are deep learning methods such as convolutional neural networks which expand the classical DSM approach by including information about the vicinity of a site. Each site is characterized by n climate and environmental covariates with a 3- dimensional array for width, length in pixels of a window centered at point p (site coordinates) and the covariates. Multi-task learning can handle the notion of depth by providing prediction layer by layer. More research and investigation should be done to explore how to apply these deep learning approaches can be challenging competitions of our hierarchical non linear structure model with latent variables. The possibility of extrapolating the statistical model is very useful since optimizing the sampling design takes a lot of time and it is also expensive (acquisition of data and processing samples in the laboratory).

- **Bayesian selection approaches can help to better understand the outputs of the mechanistic model for soil carbon dynamics**

The coefficient of diffusion, which underlines soil bioturbation, and the advection coefficient, which is related to lessivation, are treated as constants in the mechanistic models developed for soil carbon dynamics. However, in reality, these coefficients are not constant and vary with depth. A challenge will be to transform these constant coefficients to functions that decrease exponentially with depth. The explored Bayesian selection methods explored can be used in order to define the influential climatic and environmental factors.

APPENDICES

7.1 Radiocarbon

Definition:

Radiocarbon is a radioactive carbon isotope $C^{14}C$ with a natural abundance around 10^{-12} in the atmosphere. It is the third most abundant isotope of carbon, besides ^{12}C (99.9% of natural carbon) and ^{13}C (0.1%), both of which are stable. It is produced by interactions between cosmic radiation and nitrogen atoms. Oxidized to CO then CO_2 molecules, ^{14}C is then quickly mixed within the atmosphere. Age measurements are possible because ^{14}C becomes part of all organic and inorganic carbon compfounds and a steady state between the uptake (photosynthesis or food) and the decay of ^{14}C exists as long as the organism is alive (Libby et al., 1949). After death, the only remaining process is decay (^{14}C decays into ^{14}N by beta decay). Measurement of the remaining ^{14}C atoms gives a measurement of the time that elapsed since the steady-state was broken. In 1946, Libby evaluated the half-life of radiocarbon at 5568±30yrs, later revised to 5720±7yrs. He signed the birth certificate of radiocarbon dating and the method can now be applied to reliably date materials as old as 50,000-55,000 years.

$$t = \frac{T_{1/2}}{ln2} * ln\left(\frac{A}{A_0}\right) \tag{7.1}$$

with A, ^{14}C activity at the time of dating, A_0, initial 14C activity at the time t_0 (deposition, death), $T_{1/2}$ = 5568 yr (Libby half-life).

Radiocarbon cycle issues:

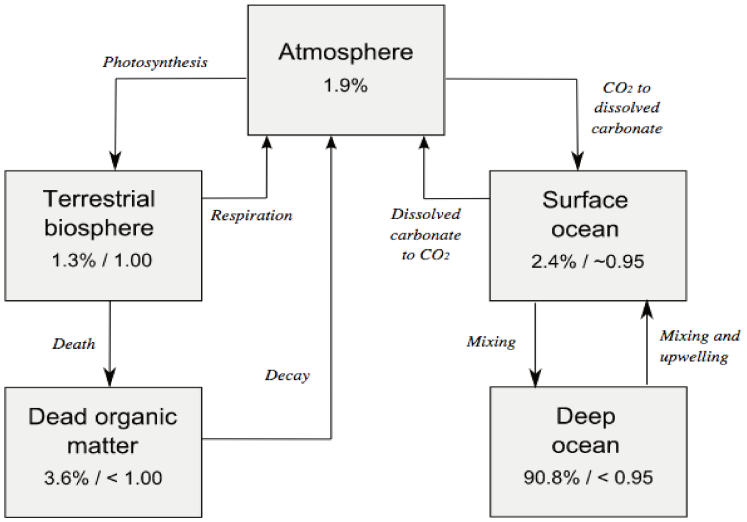


Figure 7.1: Percentages show the fraction of the total carbon reservoir of each type. Numbers after the slash show the ratio of ^{14}C to ^{12}C as a fraction of the atmospheric ratio (Bowman and Leese, 1995).

Unfortunately life is not so simple and radiocarbon geochronology faces several issues:

1. Godwin 1962 definitively fixed the half-life of radiocarbon at 5730±30 yrs. To honor the father of ^{14}C dating, the ^{14}C community nevertheless decided to keep the 5568 yr period to report "^{14}C conventional age". This means that "conventional ^{14}C age" should be increased by about 3% to be equivalent to a real age.

2. The atmospheric ^{14}C concentration is not constant with time and has varied greatly over the last 50,000 years (Mazaud et al., 1992) (Hughen et al., 2004). This is the result of changes in the balance between ^{14}C production and ^{14}C sinks. ^{14}C production results from interaction between cosmic rays and nitrogen nuclei. It

is thus linked to solar activity (which produces cosmic rays) and the intensity of the Earth's geomagnetic field (which repels cosmic rays). Both parameters have changed greatly with time and thus ^{14}C production has varied greatly as well, by a factor of 2 in the last 50,000-year period (see Fig. 7.2).

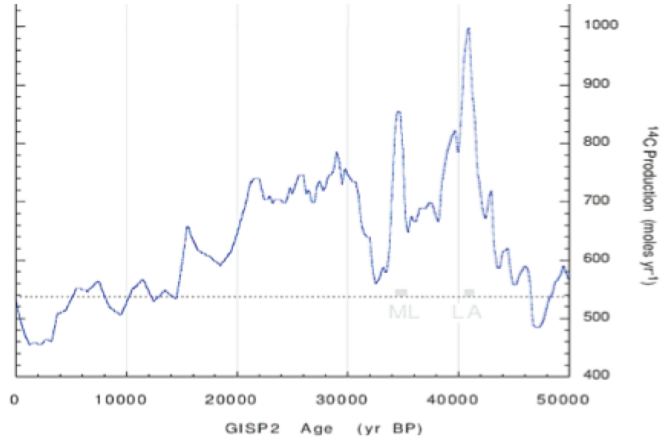


Figure 7.2: The natural variation of radiocarbon production in the last 50,000-year period (Hughen et al., 2004)

On the other hand, sinks of ^{14}C consist of carbon trapped in vegetation, soil and/or the ocean (Figure 7.1). All these reservoirs have changed in size and exchange capacity with time. For example, during glacial times (50-12 kyears before today), oceans were smaller (water was trapped in ice-sheets), much cooler and ocean currents were much weaker than today. All these aspects induce a decrease in the ocean-atmosphere exchange capacity and thus impact the ^{14}C atmospheric concentration. Taken all together, these parameters that impact ^{14}C atmospheric concentration are highly variable and too complex to model. The solution to translate ^{14}C activity into a real age was found in the late 1970s – early 1980s with the constitution of the first "calibration curves" that account for all the changes in atmospheric concentration, enabling a real age to be determined from ^{14}C measurement. The first unanimously acknowledged calibration curve was published in 1986 by Stuiver and Reimer and covered only the last 10 kyrs. The last one (Calib13 – (Reimer et al., 2013)) covers the whole period of the last 50,000 years. Calibration curves are established by comparing the absolute age of a sample (tree-ring counting, independent dating methods) and its ^{14}C content (see Fig. 7.3).

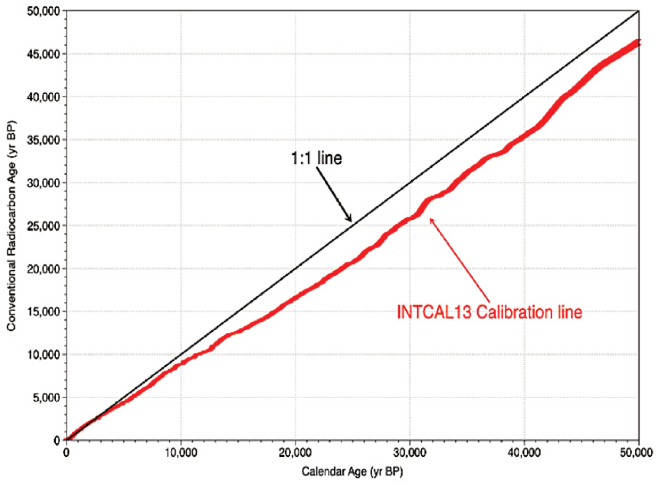


Figure 7.3: IntCal 13 by Reimer et al.(the shift between the 1:1 line and the trend of the calibration is due to the false half-time, the variations around the trend result from changes in production and changes in the ^{14}C cycle with time.

3. The difference in size and relative ratio of the carbon reservoirs (vegetation, ocean, ...) between Northern and Southern Hemispheres also impacts the ^{14}C cycle and the corresponding atmospheric ^{14}C concentration, leading to a higher content in the North than in the South.

4. Beside this natural variation of ^{14}C atmospheric content, anthropogenic activities also have an impact on the ^{14}C cycle. In the 1960s, Hans Suess showed that extensive fossil combustion decreased ^{14}C by 3% in the atmosphere (Suess, 1955), since the fuels are so ancient that they contain no ^{14}C at all (see Fig.7.4).

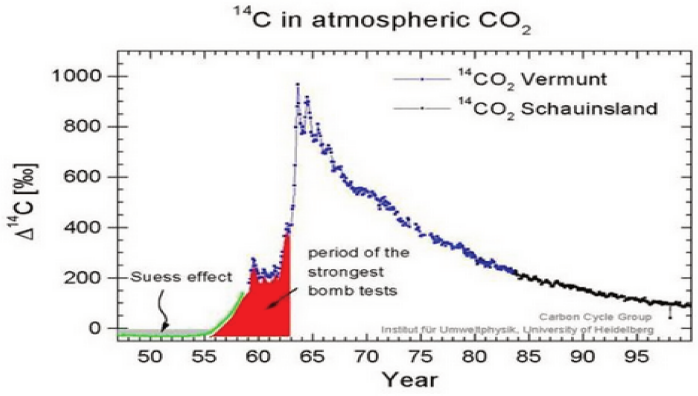


Figure 7.4: Development the atmospheric radiocarbon in the Northern Hemisphere in the last 50 years. Data before 1959 were derived from tree rings (Stuiver and Quay, 1981). From 1959 to 1983 measurements were performed at the Alpine site Vermunt; data from 1984 onwards are from the Schauinsland in the Black Forest (http://slideplayer.fr/slide/12319388/)

In the 1950s with a pejoration in the early 1960s, a much larger variation was due to the massive introduction of ^{14}C from above-ground nuclear testing (see Fig. 7.5). This massive introduction of ^{14}C did not immediately spread to all terrestrial carbon reservoirs, as shown by the double amount of natural ^{14}C concentration recorded in 1965 specially in the Northern Hemisphere atmosphere. The Partial Nuclear Test Ban Treaty was signed on 10 October 1963, prohibiting all test detonations of nuclear weapons except for those conducted underground. Consequently a slow decrease was observed in the following years, due to the distribution of this excess amount between the oceans, vegetation and soils. Concerning atmospheric nuclear tests, plants have undergone the same ^{14}C variations as the atmosphere and their residues were contaminated.

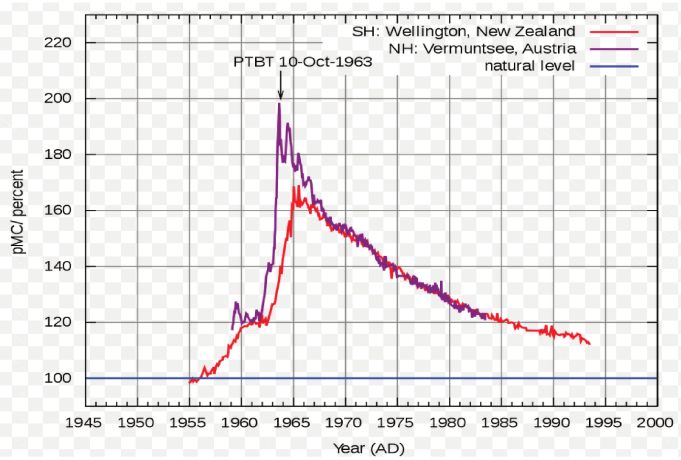


Figure 7.5: The New Zealand curve is representative of the Southern Hemisphere; the Austrian curve is representative of the Northern Hemisphere. Atmospheric nuclear weapon tests almost doubled the concentration of ^{14}C in the Northern Hemisphere (Currie, 2004)

Moreover, one can distinguish the effect of isotopic fractionation. In fact, plants fix ^{12}C more easily than ^{13}C and ^{14}C during photosynthesis. In other words, the vegetation has not the same radiocarbon concentration as the atmosphere that it grew in. To compute the isotopic fraction of $\frac{^{14}C}{^{12}C}$ we must calculate the isotopic fraction of $\frac{^{13}C}{^{12}C}$ since it is easy to measure and is not driven by time, unlike ^{14}C. The ^{14}C fractionation factor is approximately given by the square of the ^{13}C fractionation factor: $\frac{^{14}C}{^{12}C} = \left(\frac{^{13}C}{^{12}C}\right)^2$.

How to report radiocarbon activity:

Several units have been proposed in the literature to express radiocarbon activity. The choice differs from one community to another. Whereas paleoclimatologists and archeologists will prefer yr BP or cal. yr BP, geoscientists dealing with recent samples will prefer pMC, $F^{14}C$ and $\Delta^{14}C$.

To better understand the physical and the philosophical differences between the proposed units, it is necessary to come back to the physical measurement and briefly to the history of ^{14}C.

Let us assume A = $^{14}C/^{12}C$ (or $^{14}C/^{13}C$ depending on the type of measurement). The rate normalized to $\delta^{13}C$ of $-25‰$ A_N is defined as $A_N = A(1-2*(25+\delta^{13}C))/1000)$. We will note A_{SN} as the sample activity normalized to $-25‰$ and A_{ON} as the atmospheric ratio in 1950. It is retrieved thanks to two international standards: HOxI (only a few labs still use this one) and HOxII (Donahue et al., 1990). It is equal to $A_{ON} = 0.95 \times A_{HOxI}$ or $A_{ON} = 0.7459 \times R_{HOxII}$. Furthermore, the activity of the standards decreases with time and thus depends on the sampling year. This can be corrected by considering A_{abs} defined as:

$$A_{abs} = A_{ON} * exp[\lambda(y-1950)] \quad \text{with} \quad \lambda = 1/(ln(2)*5730)\text{yr}^{-1} \tag{7.2}$$

- **year BP (Before Present)**

  The first unit used is the conventional age expressed in units of years Before Present (BP). Standard practice is to use 1 January 1950 as the commencement date of the age scale. There is no particular reason for choosing the year 1950 as the reference except that it is before the bomb peak and it refers to the publication of the first radiocarbon dates. The conventional age (BP) considers the radiocarbon decay equation given in equation 7.2. Furthermore, it was calculated according to the Libby half-life of 5568 years.

  $$t = -5568/ln(2)*ln(A_{SN}/A_{ON})$$

  It was quickly recognized that yr BP were not equivalent to calendar years before 1950. It is necessary to use the calibration (e.g. IntCal13, Reimer et al. (2013)) to transform a yr BP into a real age. The real age is output as cal.yr BP. As already stated, the calibration takes into account the real period of 5730 yrs and the variations in both sources and sinks of ^{14}C with time.

- **Percent Modern Carbon, pMC**

  Because year zero is 1950, the conventional age calculated for a sample that is more recent than 1950 will be negative. This does not make sense for an age and can even cause misunderstanding. For that reason, this unit cannot be used for modern data analysis.

  In 1972, the community agreed to use pMC to replace yr BP for modern samples. This unit is considered as a percentage of the ratio of the sample's normalized activity ASN versus the absolute normalized activity A_{abs} , which corresponds to the specific activity of atmospheric carbon in the year 1950. This unit is specially used for post-bomb samples.

  $$pMC = A_{SN}/A_{abs}$$

  Problems arise when the dataset includes both old and modern samples as old samples might be reported as $pMC = A_{SN}/A_{0N}$, *i.e.* without considering the decay between the measurement year and 1950. This unit should thus be avoided for datasets such as the one we deal with here.

- **Per mil depletion**, $\Delta^{14}C$

  $\Delta^{14}C$ is a very useful way of reporting ^{14}C measurements for geochemical studies (Reimer, 2004). Under Stuiver and Polach's (1977) definition, the $\Delta^{14}C$ of a sample depends on the year in which it is measured whatever the age of the sample. For example, a sample formed in 1977 will give a different $\Delta^{14}C$ if measured today versus if it had been measured in 1977.

$$\Delta^{14}C = \left(\frac{A_{SN} e^{\lambda_c (y-x)}}{A_{abs}} - 1 \right) .1000\% \quad \text{(with age correction)}$$

  y = year of measurement;
  x = year of formation or growth;
  λ_c = (1/ln2*5730) year^{-1}.
  A_{SN}: the normalized specific radiocarbon activity for the sample;
  A_{abs}: specific activity of atmospheric carbon of year 1950, measured in 1950
  A problem might arise since $\Delta^{14}C$ is also used by the oceanography community to refer to the shift between the ^{14}C content in the ocean and the ^{14}C content in the atmosphere on the same day. As the same symbol has several meanings it might lead to confusion and one prefer to avoid it for this study.

- **Fraction Modern**, $F^{14}C$

  Reimer et al. (2004) highlight an alternative unit that does not depend on the year of measurement and is corrected for isotope fractionation. They propose as an alternative solution to use the $F^{14}C$ unit :

$$F^{14}C = A_{SN}/A_{ON} \quad A_{ON} \text{ : normalized sample activity (corrected for radioactive decay to 1950)}$$

  This unit will be adopted for our statistical analysis since it best represents the value that is closest to raw radiocarbon activity.

7.2 Bayesian modeling and inference

Bayes and Laplace

The Reverend Thomas Bayes was born in England in 1701 or 1702. The son of a Presbyterian minister from whom he took over, he was a nonconformist intellectual who seems to have lived a peaceful life as a studious bachelor, passionate about theology, natural sciences, mechanics and mathematics. His name is now associated with an elementary mathematical formula taught in any introductory course to Probability and Statistics, but Thomas Bayes would not have received the fame he now enjoys among all Bayesian statisticians if, two years after his death, a friend of his (Richard Price) had not searched his archives for Bayes' famous posthumous text - An Essay towards solving a Problem in the Doctrine of Chances - published after a reading before the Royal Society in 1763.

Here is a modern rewriting of Bayes' formula:

$$\Pr(\theta|Y) = \frac{\Pr(Y|\theta) \times \Pr(\theta)}{\Pr(Y)} \tag{7.3}$$

A few years after the death of Thomas Bayes, and independently, the French mathematician Pierre Simon de Laplace, rediscovered Bayes' formula which makes it possible to evaluate the relevance of what one believes one knows (θ) in the light of the information provided by observations (Y). Bayes, a clergyman like his father and an amateur mathematician, did not publish the one scientific paper for which he is nowadays famous, whereas Laplace resisted his father's desire for him to become a clergyman, embraced the career of –what could now be called– a professional researcher and published an impressive quantity of scientific papers in several domains. He made many advanced applications, as diverse as celestial mechanics by predicting the motion of planets and their satellites through the first statistical analysis of astronomical data, or demography by working from samples of the number of births, marriages and deaths, or reliability by studying testimony in court, and so on.

In the centuries that followed, interest in the statistical learning equation 7.3 seemed somehow to fall dormant in the academic world, despite many remarkable applications in engineering and operations research. Does history remember that Poincaré's plea for Dreyfus's innocence was based on Bayes' formula? That it was the key that allowed Alan Turing and his colleagues to break through the encrypted messages of the Enigma machine used by Nazi Germany in its military communications, giving the allies a decisive advantage? The book by McGrayne (2011) traces back through centuries the stunning story of Bayes'rule and its important, although almost secret, operational applications.

Bayesian reasoning

As a very striking example, let's skip two hundred and fifty years from Bayes and Laplace into our present world and consider, as Raftery et al. (2017) did, the following hypothesis θ : The global temperature increase of our planet will be less than two degrees by the end of the century. Bayes' formula shows that, if you wish to infer the probability of this hypothesis θ, knowing that it is some set of observations Y that has occurred, you only need to be able to quantify the probability of this collection of observations Y under hypothesis θ (and, in this sense, in general, the reasoning of physicists goes very easily from cause to observable consequences), to weight it by the chance that you agree with cause θ before observations Y are revealed, then to re-normalize this product by the probability $\Pr(Y)$.

We have here all the terms of eq7.3. The future emission Y of greenhouse gas CO_2 can be viewed as the product of the Gross Domestic Product per capita (GDP), the carbon intensity (CO_2 emission per unit of GDP) and the world population : these three driving components can be calibrated on the last 30 years of data and assessed for UN population predictions up to 2100 . The authors also modeled the links from θ to Y to calculate $\Pr(Y|\theta)$ with mild hypotheses: a decreasing trend for carbon intensity was considered and the GDP could not overcome

a world technology frontier toward which developing countries may converge. Taking into account the initial probabilities ($Pr(\theta)$) of each possible future climate change scenario informed by the Intergovernmental Panel on Climate Change specialists, the simple equation 7.3 tells us that, $\Pr(\theta|Y)$, the likely range of global temperature increase is $2.0-4.5$ (i.e. the 90% credible interval) with median 3.2 and there is only a tiny 5% chance that it will be less than $2°C$! This scientific report is represented in the form of a probabilistic judgment (Fig 7.6) that reflects the authors'uncertainty about CO_2 emissions in 2100.

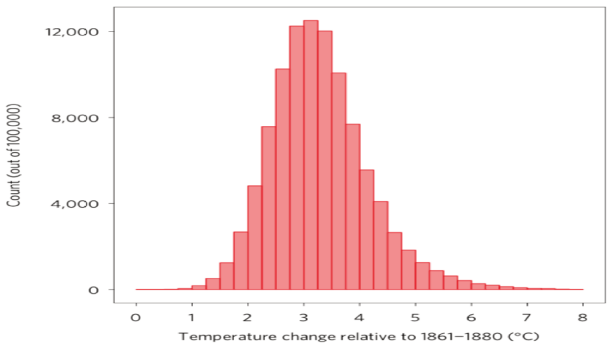


Figure 7.6: Posterior predictive distribution of the global mean temperature increase relative to 1861–1880 in Celsius degrees.
This histogram is extracted from Fig. 3 of Raftery et al. (2017)

Bayesian reasoning differs from so-called classical statistical analysis. Bayes'rule can be understood as a very generic statistical learning mechanism as summarized by Fig 7.7. The information produced by the sample Y is used to update prior beliefs about the unknown of interest (denoted θ hereafter). If more data become available, the same updating engine can be put to work sequentially.

Bayes' rule says how knowledge should change when collecting data, but does not state the starting point of this probabilistic updating machinery. The *a priori* information may not be based solely on observations, but it can also come from other sources (previous statistical analysis, experts' knowledge, etc.). The *a priori* information, which is sometimes vague, is expressed in terms of a probability distribution. This distribution is to be interpreted as a rational bet over the probable values of the unknowns. The mathematical coherence of this subjective *a priori* is enforced by the rules of probability. Compared to a classical analysis, the task of the Bayesian statistician is two-fold :

- modeling the occurrence of the observed data Y,
- and modeling the *a priori* distribution of the unknown θ.

From this *joint* model , Bayes' theorem is used in order to identify the *a posteriori* distributions of the unknown given the data. This distribution represents the actualization of our *a priori* beliefs with regard to the new information conveyed in the observations and promoted by the model (see Fig.7.7):

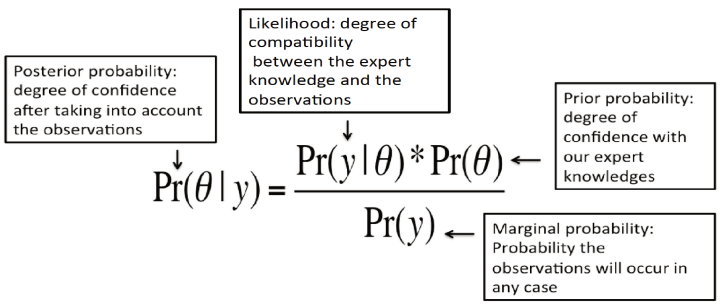


Figure 7.7: General statement of Bayes theorem.

Decisions and inferences for model parameters are based on the posterior distribution. For details on the theory of Bayesian inference, see Berger (2013) and Robert (1992). The diagram in Fig.7.8 shows the basic steps of Bayesian reasoning that lead to the *a posterior* inference

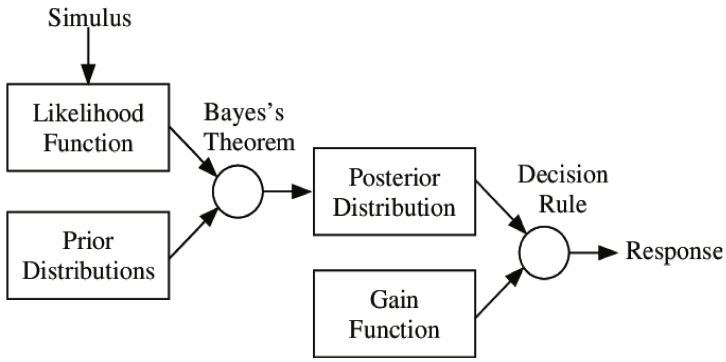


Figure 7.8: Illustration of the fundamental steps of Bayesian analysis (Healy and Proctor, 2003).

A powerful illustration of the value of the Bayesian rationale is exemplified by the search planning by Stone et al. (2011) for the wreckage of the Air France Flight 447 that disappeared , with 228 passengers and crew aboard, over the South Atlantic during a night flight from Rio de Janeiro to Paris in June 2009. In July 2010, the French Bureau d'Enquêtes et d'Analyses missioned Metron (a scientific consulting company dedicated to solving challenging problems with advanced mathematical methods) to review the three previous unfruitful searches (one acoustic search for the locator's beacons and two sonar scannings of the ocean floor) , and produce posterior probability maps for the location of the wreckage such as fig 7.9 . Using careful and methodical consideration of all the previous data available, with associated uncertainties, an analytic assessment of the highest likelihood areas for future search efforts was issued through the Bayesian formula 7.3. After one week of search, the wreckage was finally located in a high probability area of the map in April 2011...

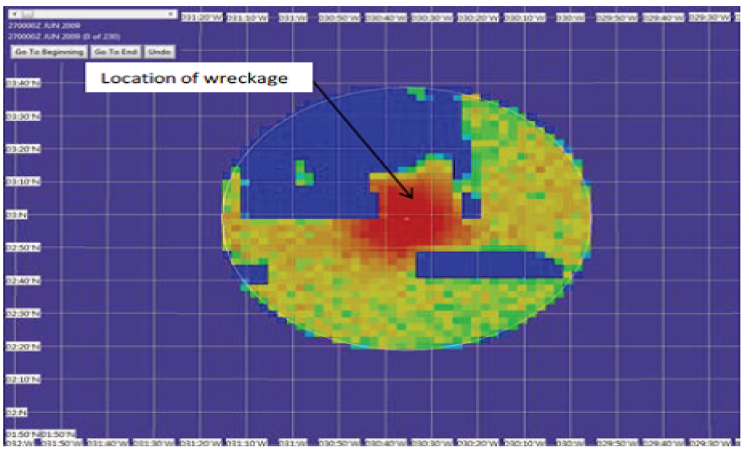


Figure 7.9: Illustration of the Bayesian analysis for the wreckage search of Air France Flight 447 (High likelihood areas given in Fig 13 from Stone et al. (2011))

Hierarchical Bayesian Models

Modeling jointly Y and θ often appears not so straightforward and it is usually necessary to go through an intermediate step to describe the process before the observations. Hierarchical Bayesian Modeling (HBM) is a specific Bayesian modeling strategy with the capacity to cope with high-dimensional complex models (Berliner,

2003; Wikle, 2003; Clark and Gelfand, 2006; Cressie et al., 2009).

HBM works through the conditional decomposition of high-dimension problems into a series of probabilistically linked simpler substructures. HBM makes it possible to exploit diverse sources of information to derive inferences from large numbers of latent variables and parameters that describe complex relationships while keeping as close as possible to the basic phenomena.

The three basic layers of hierarchical statistical models are as follows:

1. A parameter (often called θ) level identifying the fixed quantities that would be sufficient, were they known, to mimic the behavior of the system and to produce new data statistically similar to the ones already collected.
2. A latent process level Z depicting the various hidden mechanisms (given the parameters θ) that make sense of the data;
3. A data level that specifies the probability distribution of the observables at hand (Y) given the parameters (θ) and the underlying processes (Z) ;

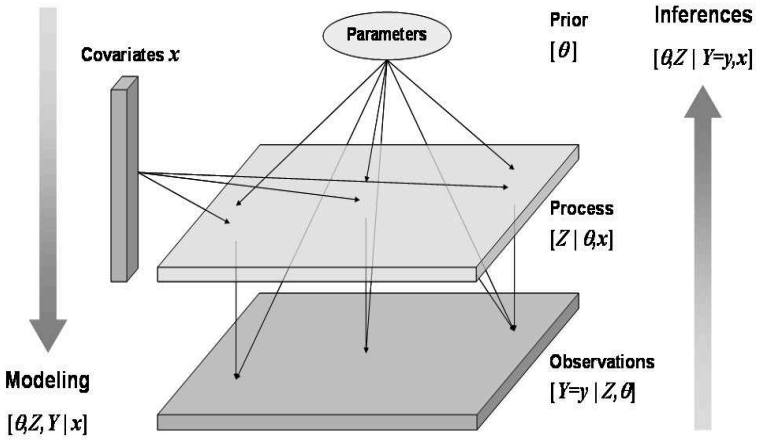


Figure 7.10: Hierarchical modeling strategy. Factorization of the complexity and Bayesian inferences. This representation is extracted from Fig. 1.12 of Parent and Rivot (2012) with bracket notations for probability distributions.

HBM stands out as an approach that can accommodate complex systems in a fully consistent framework and can represent a much broader class of models than the classical statistical methods from ready-to-use toolboxes that can be found in the frequentist literature. Eq 7.3 still applies by changing the notations and considering the block of unknowns (θ, Z) as the term that was formerly named θ.

Computational Bayes

Since the turn of the century, Bayes' formula has also revolutionized (Brooks, 2003) the scientist's toolbox: simulation methods, also known as Monte Carlo techniques in statistical jargon, made possible by the successive waves of (personal) computer advances , are effectively helping to disentangle the complex networks of causes that appear in the scientific challenges at the beginning of our third millennium.

Obtaining the posterior distribution of the unknowns (the term $Pr(\theta|Y)$ in equation 7.3) is a more difficult task than it appears at first sight : the denominator of the formula ((the term $Pr(Y) = \int_\theta Pr(\theta|Y) \times Pr(\theta) d\theta$) is an integral which needs to be performed on the support of the unknowns but θ may vary over a high dimensional space!

Brute force numerical integration does not work beyond dimension 3 , and for long, only reduced dimension models -not to say baby models - were under study in the Bayesian paradigm. An alternative approach was models for which the association between the prior $Pr(\theta)$ and the likelihood $Pr(Y|\theta)$ appears in closed form; these are the so-called conjugate models. Working in such restrictive ad-hoc families of models curbed the impetus of the lay Bayesian modeler from the twentieth century. However algorithms to tackle the problem of generating random draws from equation 7.3 without computing its denominator were already known by the end of World War 2 because a lot of effort had been devoted to their development in the Manhattan project at Los Alamos where there were computers at the time, but computer power in ordinary research laboratories was not yet sufficient to allow for their routine implementation. It is only since the turn of the century (around 1990) that Markov Chain Monte Carlo (MCMC) algorithms, with the advent of personal computers, became easily available worldwide for scientists. For details on the modern theory of Bayesian inference, see Brooks et al. (2011) and Neal (2011).

Monte Carlo integration

Bayesian learning requires evaluating the expectation of functions of the unknowns θ with respect to the posterior distribution. Nowadays, Monte Carlo algorithms avoid the explicit computation of the Bayes formula denominator since they only need to know the distribution from which to sample, up to a constant. To focus on the technical problem , we will now write $\pi(\theta)$ for the posterior probability density (because from now on we consider that the data are given once and for all and need not be recalled in $Pr(\theta|Y)$ and review some techniques to evaluate the essential characteristics of $\pi(\theta)$ such as

$$E_f = \int_\theta \pi(\theta) f(\theta) d\theta$$

for any measurable function f when $\pi(\theta)$ is known up to a constant. For instance, if the test function $f(\theta) = \theta$ one would obtain E_f as the posterior mean, and more generally the moments of order p (when they exist) by taking $f(\theta) = \theta^p$. Such an expectation of f can be approximated by the Monte Carlo method, using a G-sample $(\theta^{(1)}, \ldots, \theta^{(G)})$ of independent values generated from π

$$E_f \approx \frac{1}{G} \sum_{g=1}^{G} f(\theta^{(g)}) d\theta$$

Unfortunately, a method to directly draw an independent sample of $\theta^{(g)}$ from π is generally not available for π which is often a multivariate sophisticated distribution (it is the unnormalized product of the likelihood and the prior) . But the approximation will still work when the $\theta^{(g)}$ are dependent, as long as the dependence is not too stringent, such as the one given by a Markov chain (with specific properties easy to enforce in the case of homogeneous Markov chains). An ergodic Markov chain, with π as a (unique) invariant distribution, will converge to π from any initial distribution of states given to the chain. MCMC are ergodic Markov chains designed to stochastically visit the support of θ , quickly disregarding the initial values of the chain (the so called burn-in period) and moving rapidly – i.e without too many correlations between the successive steps of the algorithm – so as to reach the equilibrium distribution π and make it possible to have access to the posterior distribution by means of the random sample of replicates $\{\theta^{(g)}\}_{g=1:G}$ (Brooks et al. (2011); Gelman et al. (2013a); Roberts (1996); Robert (2004)).

The free software WinBUGS (Lunn et al., 2000; Congdon, 2001; Spiegelhalter et al., 2003; Ntzoufras, 2011) is a tool of choice for Bayesian inference when the posterior distributions contain high-dimensional latent system states and parameters, which is the case of HBMs. OpenBUGS is the open source variant of WinBUGS which offers greater flexibility and extensibility. JAGS is a program developed by Martyn Plummer that relies on the same modeling grammar and language as WinBUGS, but its main advantages are its speed (since it is written in C++ instead of Component Pascal) and its platform independence (with regards to Windows systems) (Plummer et al., 2003). Many step-by-step Bayesian initiation course tutorials have been developed using OpenBUGS/JAGS and their extensions (GeoBUGS, PkBUGS, etc.). King et al. (2010) is of special interest for the practitioner because, for each case study, they provide their inference routines both in R (as a stand-alone program) and in WinBUGS.

NIMBLE (Numerical Inference for Statistical Models using Bayesian and Likelihood Estimation) is an interesting R package developed by Perry De Valpine, Christopher Paciorek, Duncan Temple, and Daniel Turek whose first version was published in June 2014 . The package (de Valpine et al., 2017) is designed to deal with hierarchical models and the many problems they raise. The NIMBLE creators try to fill the gap between an abundant literature on these questions but a limited software offer that does not allow scientists to write their own blocks of the inference routine. The idea underpinning NIMBLE is to allow both a flexible model specification and a programming system that adapts to the model structure. Some frequentist authors even resort to Bayesian computation as an ultimate tool in the hand of the analyst to find solutions to problems not solved by other conventional analyses, see for example Chapter 17 of Zuur et al. (2009) or the many papers on data cloning by Lele et al. (2007).

Gibbs sampling

Gibbs sampling is the simplest MCMC method, widely used for Bayesian inference following its introduction by (Geman and Geman, 1993) and (Gelfand and Smith, 1990). It is applicable when the multi-dimensional parameter $\theta = (\theta_1, \ldots, \theta_j, \ldots, \theta_p)$ is such that one can easily sample from each of the full conditionals of the posterior distribution. The full conditional for one component θ_j of θ is the conditional distribution (with respect to π) of this j^{th} component given values for all the other components of θ. The Gibbs sampler generates a MCMC algorithm [1] by stacking stochastic transitions from $\theta^{(g)}$ to $\theta^{(g+1)}$ relying on the full conditionals as follows:

- Pick $\theta_1^{(g+1)}$ from the distribution of θ_1 given $\theta_2^{(g)}, \ldots, \theta_p^{(g)}$
- Pick $\theta_2^{(g+1)}$ from the distribution of θ_2 given $\theta_1^{(g+1)}, \theta_3^{(g)}, \ldots, \theta_p^{(g)}$
- ...
- Pick $\theta_j^{(g+1)}$ from the distribution of θ_j given $\theta_1^{(g+1)}, \theta_2^{(g+1)}, \ldots, \theta_{j-1}^{(g+1)}, \theta_{j+1}^{(g)} \ldots, \theta_p^{(g)}$
- ...
- Pick $\theta_p^{(g+1)}$ from the distribution of θ_p given $\theta_1^{(g+1)}, \theta_2^{(g+1)}, \ldots, \theta_{p-1}^{(g)}$

The Gibbs sampler has the favor of many practitioners since no tuning is necessary to run the MCMC algorithm. It grounds the WinBUGS expert system (BUGS meaning Bayes Using Gibbs Sampler) and the JAGS (Just Another Gibbs Sampler) package. The Gibbs sampler takes advantage of the conditional independence structures encountered in the direct acyclic graph nodes of HBM to implement more efficient sampling in the parameter space (Lauritzen and Spiegelhalter, 1988): many conditional distributions are of standard forms for which efficient sampling procedures are readily available. However Gibbs sampling may be rather slow since drawing in conditionals generally yields only small moves in the parameter space and this drawback worsens with the parameters' dimension.

Metropolis

The Metropolis algorithm was first presented in the seminal paper of Geman and Geman (1993) but statisticians had to await the development of personal computers for a generalized diffusion of this sampling routine, able to generate random draws from almost any distribution known up to a constant. Conversely to the Gibbs sampler, the Metropolis algorithm needs an auxiliary distribution (or jump function $J(\theta' \rightarrow \theta)$) to make proposals that will be accepted or rejected. In the latter case, the new value is a replication of the previous state of the algorithm. The basic Metropolis generates an MCMC algorithm [1] by iterating singular probabilistic transitions from $\theta^{(g)}$ to $\theta^{(g+1)}$ as follows:

- Generate a candidate $\theta^\star$ from $J(\theta^{(g)} \rightarrow \theta^\star)$
- Accept the candidate $\theta^\star$ with probability $min(1, \frac{\pi(\theta^\star)}{\pi(\theta^{(g)})})$ [2]

[1] to be checked for ergodicity although obtained in most cases

[2] if $\pi(\theta^\star) > \pi(\theta^{(g)})$ the candidate is accepted anyway, else the ratio $\frac{\pi(\theta^\star)}{\pi(\theta^{(g)})}$ is compared to a uniform random draw

- Upon acceptance of the candidate , let $\theta^{(g)} = \theta^{\star}$, else replicate the current value by setting $\theta^{(g+1)} = \theta^{(g)}$

If the proposal distribution does not fulfill the symmetry condition $J(\theta' \rightarrow \theta) = J(\theta \rightarrow \theta')$, the ratio of acceptance has to be adapted. Many choices can be made for the proposal distribution , the simplest one being the Gaussian distribution centered on $\theta^{(g)}$ with a variance to be carefully tuned so as to monitor the ratio of acceptance of the candidate. Tuning can be rather challenging, however, following the *Goldilocks principle*, according to Rosenthal (2014). If the candidate is too rarely accepted (because the average jump to make a proposal –controlled by the variance of the proposal distribution– is too big), the chain remains stuck for long periods , but on the contrary, if the candidate is too frequently accepted (because the variance is too small) the chain wobbles within a small area of the parameter space). In both cases, the tuning of the jump function is bad, resulting in a slow convergence of the MCMC targeting the posterior distribution $\pi(.)$.

Beyond Gibbs and Metropolis

There is a huge continuous development of stochastic algorithms devoted to improving the tools needed to perform Bayesian inference. Many paths for future developments can be anticipated.

- There have been many attempts to develop adaptive versions of MCMC (Rosenthal, 2014; Atchadé et al., 2010) but automatically tuning the algorithm remains a challenging task, because convergence properties of non homogeneous Markov Chains are difficult to establish . Relying on weighted independent draws such as the one obtained through the family of importance sampling techniques seems a more promising avenue of research (Liu et al., 2001). To design efficient importance sampling algorithms, the auxiliary proposal distribution should be chosen as close as possible to the posterior distribution but as the posterior distribution is unknown, choosing such a proposal is a tricky crucial task (Gelman et al., 1996; Roberts et al., 1997). The adaptive multiple importance sampling algorithm of Cornuet et al. (2012) is a good example of the efficiency of Population Monte Carlo methods (Cappé et al., 2004). It consists in iteratively generating parameters under an adaptive proposal distribution and assigning weights to the parameter replicates. The AMIS algorithm sequentially tunes the coefficients of the proposal distribution (selected from a parametric family of distributions, generally the Gaussian one) and recomputes the weights of the cumulated posterior sample at the end of each iteration.

- Hamiltonian Monte Carlo algorithm –originally known as hybrid Monte Carlo (Duane et al., 1987)–, is a most remarkable Markov chain Monte Carlo method for reducing the correlation between successive sampled states by using properties from Hamiltonian dynamics (Neal, 2011). It allows larger moves between states at the cost of doubling the dimension of the state space and being able to efficiently compute the gradient of the logposterior density. This causes the algorithm to converge more quickly to the targeted posterior probability distribution. Carpenter et al. (2017) developed the No-U-Turn sampler, an adaptive form of Hamiltonian Monte Carlo sampling that is encoded in the package Stan. Stan, now available in a stabilized version, provides useful modern tools for Bayesian inference for continuous-variable models that are used in a wide range of application domains, e.g. in ecology (Monnahan et al., 2017).

- Variational Bayes methods (Beal et al., 2003) drop the idea of targeting the exact posterior distribution but search for an approximate (with respect to a Kullback-Leibler divergence) solution belonging to some convenient multivariate parametric family, by alternating quick optimization and expectation steps. A particularly successful variational approximation in pattern recognition and machine learning is the factorized form (Jordan et al., 1999a,b). The idea for such a factorized approximation stems from theoretical physics where it is called mean field theory (Parisi, 1988). If one wishes nevertheless to get the exact posterior distribution, Donnet and Robin (2017) proposed a rather straightforward implementation of a bridge sampling scheme (Gelman and Meng, 1998) starting from a variational approximation of the posterior distribution and targeting the true one.

Going Bayesian

The application of Bayesian concepts and methods was for long poorly developed because of the complexity of posterior calculations and the absence of closed form solutions. Advanced numerical computation for the implementation of Bayesian methods (Monte Carlo Markov Chain simulation methods: Gibbs Sampler, Metropolis Hastings, Hamiltonian MCMC, Population Monte Carlo, etc.) nowadays allow us to overcome this technical obstacle . This is reflected in a marked upsurge of Bayesian methods in all statistical applications (Hoff, 2009). When searching for the term " Bayesian" in PubMed, one can observe an exponential growth in the number of publications, which has doubled about every five years since 1963! Bayesian models and algorithms are now widely applied in all domains of engineering and machine learning (Barber, 2012), as well as neural and psychological models. There even is a theory of the Bayesian brain (Doya et al., 2007), following the seminal work of Dehaene (2011).

Of course, the merits of Bayesian reasoning are grounded on philosophical arguments regarding uncertainty such as the ones detailed in Kadane (2011) and Lindley (2013), but people tend to attach more importance to the demonstration of Bayesian effectiveness in practical contexts. Among the many advantages of Bayesian statistical inference that explains its seemingly inexorable success, the following ones are often pointed out:

- Bayesian learning is simply performed by applying the rules of probability and data assimilation is nothing but updating the scientist's probability distribution over all the unknown quantities. Bayesian inference offers a direct probabilistic interpretation for credible intervals and tests, as well as fruitful insight into frequentist estimators and statistical decision rules.
- The uncertainties on all unknowns are taken into account in the scientific reporting of the study and the predictive applications of a model, often delivering a less optimistic point of view than their frequentist counterparts.
- The missing values can be easily imputed using Bayesian MCMC simulation methods. The analysis is often more robust to outliers, by using more flexible distributions.
- For small sample surveys, Bayesian analysis can help to increase precision by bringing neighboring information or prior expertise into the analysis. For large sample surveys, central limit theorems often make it possible to obtain the frequentist classical results as asymptotic limits of Bayesian ones .

7.2.1 Bayesian model checking and Bayesian model comparison

A- Bayesian model checking

After specifying the prior distributions based on the expert advice, constructing a reasonable model to explain the data and computing the posterior distributions of the unknowns based on the MCMC simulation methods mentioned previously, a last remaining step is to check the adequacy of the model fit.

A crucial point to check is the robustness of the results to the choice of prior distributions by a **sensitivity analysis** . It is important to examine how the posterior distributions change when another prior distribution is used instead of the present one (Gelman et al., 2013a). In fact, the prior distribution should represent our probabilistic judgment about the values that the parameter to be estimate may have. However, it is sometimes very difficult to come up with precise distributions to be used as prior. Thus, the analyst would like to test all possible combinations of prior distributions and likelihood functions selected from classes of priors and likelihoods considered empirically plausible. The Bayesian model is considered to be robust if the posterior distributions seem invariable with the different priors. If the posterior results differ substantially, the analyst must either make the prior informative using expert advice or build several prior + likelihood assemblies and specify the set of modeling assumptions considered

for each case. For example, in the case of Bayesian selection methods, it is useful to study the effect of fixing the prior inclusion probability "p" to 0.5 or proposing a prior distribution as a Beta(2,2) or a Uniform(0,1) distribution. A common choice among Bayesian selection approaches is to set the prior inclusion probability to 0.5 as this choice makes all models equiprobable but on the other hand, it favors models where about half of the variables are selected. This choice, for example, is problematic when a large number of explanatory covariates are included in the model. Another point is to investigate if **the inferences from the model make sense**. Sometimes, the expert does not intervene to specify prior distributions for reasons of convenience or objectivity or time, thus the posterior results may not reveal what they expect. For example, for the soil scientist, atmospheric radiocarbon is an influential predictor of topsoil radiocarbon. However, the application of a Bayesian selection model does not detect atmospheric radiocarbon as a significant predictor. This result may not make sense for soil scientists but there may be some explanations for this lack of significance. For example, on the one hand, the natural variation in atmospheric radiocarbon is not constant with time and on the other hand, it is marked by an artificial increase from above-ground nuclear testing from about 1950 until 1963. Consequently, it is clear that the database is not representative of the actual variation in atmospheric ^{14}C (the majority of data were sampled after 1963). Furthermore, the prior inclusion probability is fixed at 0.5 which means that the same probability is given to select or remove the predictor. It is not a good idea to set non informative priors and rely on Bayesian inference as a data analysis approach. In such a case, to achieve a more coherent statistical selection, the expert should first express his own beliefs on the prior inclusion probability for atmospheric ^{14}C , setting the prior inclusion of this predictor to a value (higher than 0.5) that corresponds to his own belief, and see if the updated posterior distribution after data assimilation goes against his prior judgment .

Furthermore, investigations should be carried out to test whether **the model is consistent with the data**. In fact, a model is consistent if replicated data generated from the model look similar to the observed data. An important discrepancy can highlight misfitting of the model. The discrepancy between model and data can be calculated based on the tail-area probability or the so-called Bayesian p-value applied to the test quantities $T(y,\theta)$ and $T(y_{rep},\theta)$, where y represents the observed data, y_{rep} represents the replicated data and θ the unknown parameters that we seek to estimate (Gelman et al., 2013a). The Bayesian p-value is defined as the probability that the replicated data will exceed and be more extreme than the observed data:

$$p_B = Pr(T(y_{rep},\theta) \geq T(y,\theta)|y)$$

If the model is true or close to true, the posterior predictive p-value will almost certainly be very close to 0.5 (Gelman et al., 2013b). However, a clear guide-line on how best to interpret the p-value is not available. In fact, the posterior predictive posterior check remains one of the most misunderstood and confusing concepts in statistics. Gelman stressed the fact that a p-value – say equal to 0.2 – is not, and should not be interpreted as a claim that the model is “true”; rather, it should be interpreted as a statement that the model (probabilistically speaking) fits one particular aspect of the data (summed by the T statistics).

B- Bayesian model comparison

The usefulness of the model is evaluated by its ability to provide a good data fitting and prediction about the future. There exist several Bayesian criteria to compare different models in competition. We distinguished among them:

- **The Deviance Information Criterion (DIC)** is a measure of predictive accuracy. This criterion is defined as follows:

$$DIC = 2\hat{D}_{avg}(y) - D_{\hat{\theta}}(y)$$

$$\hat{D}_{avg}(y) = \frac{1}{L}\sum_{l=1}^{L} D(y,\theta^l) \quad \text{and} \quad D_{\hat{\theta}}(y) = D(y,\hat{\theta}(y))$$

where $D(y,\theta) = -2log(P(y|\theta))$ and L is the number of iterations. $\hat{\theta}$ is fixed to the posterior mean of θ under the posterior distribution. The estimated average $\hat{D}_{avg}$ is a better summary of discrepancy than $D_{\hat{\theta}}$ of the point estimate since it takes into account the model uncertainties. The DIC is a popular criterion among the community of Bayesian practitioners, used to compare models through its easy implementation in the graphical modeling package BUGS/JAGS/Stan. The model with the smallest DIC is considered to be the model that would best predict a replicated dataset with the same structure as the currently observed one (for more details see the paper by Spiegelhalter et al. (2002)). However for most Bayesian theorists, this criterion is not theoretically grounded since it is defined using a pointwise estimate (the posterior mean estimation $\hat{\theta}(y)$) without any probabilistic justification.

- **Widely Applicable Information Criterion (WAIC)** is seen as an improvement of the DIC. As mentioned previously, the DIC is considered as non fully Bayesian and does not work for singular models. The WAIC claims to be fully Bayesian since it is based on the entire posterior distribution. WAIC can be calculated as:

$$WAIC = \frac{1}{n}\sum_{i=1}^{n} log(P(y_i|y)) - \frac{V}{n}$$

where, $V = \sum_{i=1}^{n} Var(log(P(y_i|y))$. The package "loo" developed in R allows fast computation of the WAIC. The function compares 2 models by estimating the difference between prediction errors. The difference will be positive if the expected predictive accuracy for the second model is higher (for more details, see Watanabe (2010); Vehtari et al. (2015)).

- **Bayes Factor (BF)** can be interpreted as how much the data favor model M_1 over M_2. It is equal to the ratio of the evidence (prior mean of the likelihood) of one particular hypothesis to the evidence of another:

$$BF = \frac{P(y|M_1)}{P(y|M_2)} = \frac{\int_\theta P(y|\theta,M_1)P(\theta|M_1)}{\int_\theta P(y|\theta,M_2)P(\theta|M_2)}$$

Its most common interpretation is the one -—first proposed by Jeffreys (1998) and slightly modified by Lee et al. (2014)– given in Table 7.1:

Bayes factor	Evidence category
$\geq$ 100	Extreme evidence for M_1
30-100	Very strong evidence for M_1
10-30	Strong evidence for M_1
3-10	Moderate evidence for M_1
1-3	Anecdotal evidence for M_1
1	No evidence
1/3-1	Anecdotal evidence for M_2
1/10- 1/3	Moderate evidence for M_2
1/30-1/10	Strong evidence for M_2
1/100-1/30	Very strong evidence for M_2
$\leq$ 1/100	Extreme evidence for M_2

Table 7.1: Interpretation of Bayes Factors.

- **k-fold Cross Validation (C.V)** estimates the predictive power of the model: for example, to be able to use a Bayesian model to predict the soil carbon dynamics for sites where no measurements were taken. This procedure considers splitting the data into k groups of sub-sets. The model is adjusted on $k-1$ sub-sets "$y_{(training)}$" and a cost function will be evaluated on the remaining sub-set "$y_{(validation)}$". This procedure of training and validation is repeated k times and the k performances are averaged in order to assess the predictive

performance of the model, avoiding overfitting (Stone, 1974). k is usually taken as 5 or 10, but there is no general formal rule. As k gets larger, the difference in size between the training set and the re-sampling subsets diminishes. As this difference decreases, the bias of the technique becomes smaller. The relative error is defined as follows:

$$RE(\%) = \frac{|y_{(training)} - y_{(validation)}|}{|y_{(test,real)}|} * 100$$

When multiple models are compared, the model with smaller Relative Error (RE) is preferred to models with larger RE.

- **Coverage of Bayesian credible intervals percentage on validation sets**
  The credible interval covers a pre-specified credibility range of a given unknown. Since the validation sets are not taken into account to adjust the model, one can check that the probability that the actual observation will lie within the credible interval corresponds to its theoretical value. The percentage of coverage of the Bayesian credible interval is defined as follows:

$$p\_cov(\%) = \frac{1}{L}\left(Y_{(validation)} > born\_inf \quad \& \quad Y_{(validation)} < born\_sup\right) * 100$$

  where L represents the number of MCMC iterations, $P(Y_{(validation)} < born\_sup) = 0.975$ and $P(Y_{(validation)} < born\_inf) = 0.025$. $Y_{(validation)}$ indicates the sub sets of data which are not used to adjust the model. $born\_sup$ and $born\_inf$ correspond to the 2.5% and 97.5% quantiles.

C- Convergence of Bayesian computations

In addition, an essential point to check is the Monte Carlo convergence of the estimated unknown model parameters. Indeed, in theory, the posterior distribution will converge to the target distribution after an infinite number of MCMC iterations. Practically, the number of MCMC iterations is of course finite and does matter for computational cost and efficiency. Assessing MCMC convergence is crucial since Bayesian statistical reporting is entirely based on the posterior distribution : controlling the numerical accuracy of parameter estimation is therefore necessary to quantify uncertainty. The best way to check the convergence of the Bayesian model is to run multiple Markov chains initialized from different initial conditions. Theoretically, these Markov chains should forget their starting points and converge to the same target distribution. There exist several warning signals which indicate that convergence is not established. One can distinguish the following visual inspections:

- A low or high acceptance rate for the Metropolis Hastings simulation algorithm,
- Poor mixing when observing the trace-plots of the estimated model parameters,
- High autocorrelation between states of the Markov chain. The higher the autocorrelation in the chain, the larger the MCMC variance and the worse the approximation. The lag-t autocorrelation function of the sequence that are t steps apart is defined as (Hoff (2009), chapter 6):

$$acf_t(\theta) = \frac{\frac{1}{S-t}\sum_{s=1}^{S}(\theta_s - \bar{\theta})(\theta_{s+t} - \bar{\theta})}{\frac{1}{S-1}\sum_{s=1}^{S}(\theta_s - \bar{\theta})^2}$$

  A high value of the acf indicates that the MCMC is only slowly moving around the parameter space and may take a long time to explore the parameter space adequately.

- Suspicious tails or shapes when examining the posterior distributions for unknown parameters.

There also exist several formal tests to judge MCMC convergence. The most popular are : the Geweke diagnostic (Geweke, 1992), the Raftery and Lewis diagnostic (Raftery and Lewis, 1992), the Gelman and Rubin diagnostic (Gelman et al., 1992), etc. The Geweke diagnostic takes two non overlapping parts (usually the first 0.1 and last 0.5 proportions) of the Markov chain and compares the means of the two parts, using a difference of means test to see if the two parts of the chain may stem from the same distribution (as a null hypothesis). The disadvantage of this diagnostic is that it is sensitive to the specification of the spectral window. In addition, Geweke does not suggest a quantitative rule to conclude to convergence. On the other hand, the Raftery and Lewis diagnostic estimates the minimum chain length needed to estimate a percentile to some precision. One needs to select a posterior quantile of interest q, an acceptable tolerance r for this quantile and a probability s, which is the desired probability of being within $(q-r, q+r)$. Thus the minimum length is given by applying the following recipe:

$$n_{min} = \left[\phi^{-1}(\frac{s+1}{2}) \frac{\sqrt{q(1-q)}}{r} \right]^2$$

where $\phi^{-1}(.)$ is the inverse of the normal cumulative distribution function.
The convergence is based on the dependence factor I obtained by the raftery.diag() function in R. A high dependence factor (for example > 5) may be explained by bad starting values, poor mixing or a high correlation between parameters. A review paper by Cowles and Carlin (1996) pointed out some weaknesses of the Raftery and Lewis convergence diagnostic. For example, variable estimates can be produced given different initial chains starting points. Added to that, it is not realistic to impose that the convergence should be tested for every quantile of interest. Finally, the Gelman and Rubin diagnostic compares intra and inter variances of Markov chains. The implementation of the Gelman and Rubin test is available in the programs developed for MCMC simulations such as JAGS/ BUGS/ Stan, etc. The potential scale reduction factor $\hat{R}$ is defined as:

$$\hat{R} = \sqrt{\frac{\hat{Var}(\theta)}{W}}$$

where $\hat{Var}(\theta) = (1-\frac{1}{n})W + \frac{1}{n}B$ (n is the number of discarded iterations), W is the mean of the variances of each chain, defined as $W = \frac{1}{m}\sum_{j=1}^{m} s_j^2$ with s_j^2 the variance of the jth chain given by: $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta}_j)^2$. The Variance between chain is given by $B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\theta}_j - \bar{\bar{\theta}})^2$ where $\bar{\bar{\theta}} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_j$ and m is the number of total iterations.
The convergence is satisfied when the potential scale reduction factor $\hat{R}$ for each of the unknown parameters goes to 1. A high value of $\hat{R}$ (greater than 1.2, according to the recommendations of the authors) underlines the need to run the Markov chain longer so as to reach convergence to the stationary distribution.
One of the challenges of the Gelman and Rubin convergence test is to propose initial values from a starting distribution that is sufficiently dispersed with respect to the target distribution to adequately explore the parameter space ...but not too far from the target because we wish the proposals of the MCMC auxiliary distribution to be finally accepted after a reasonable time so that all the chains start evolving after initialization.

7.3 Effect of the unbalanced experimental design on the quality of the estimators

To carry out an experimental design, we have to clarify the following points:

1. The definition of the purpose and conditions of the experiment.
2. The definition of the factor or factors to study, and its or their levels.
3. The definition of individuals or, more generally, the experimental units that we propose to observe.
4. The definition of the number of observations.
5. **How to assign the different objects to the different experimental units**.

We consider a two-way analysis of variance (ANOVA) with two additive factors A and B with I and J levels respectively and without considering the interaction effect between the factor A and B. Thus, the m-th observed response when factor A is of level i and factor B is of level j is written as :

$$y_{i,j,m} = \mu_{i,j} + \varepsilon_{i,j,m} \quad \varepsilon_{i,j,m} \sim N(0, \sigma^2)$$

The mean $\mu_{i,j}$ is expressed as:

$$\mu_{i,j} = \mu + a_i + b_j$$

- $\mu_{i,j}$: is the mean of the observation for the ith level of factor A and the jth level of B.
- a_i: is the ith effect of level i within the factor A.
- b_i: is the jth effect j within the factor B.
- μ: is the grand mean of observations.
- n: is the total number of observations.

The 2-factor experimental design (A and B) is orthogonal if it verifies:

$$n_{i,j} = \frac{n_{i+} \times n_{+j}}{n} \qquad \forall i = 1, \ldots, I \quad \text{and} \quad \forall j = 1, \ldots, J$$

where n_{i+} is the total number of observations of level "i" within the factor A (respectively n_{+j} is the total number of observations of level "j" within the factor B). This definition means that if an experimental unit (observation) is selected at random, the events "observation of i level of factor A" and "observation of j level of factor B" are independent.

In multiple regression, the experimental design is orthogonal if the explanatory variables are not correlated. The orthogonality of experimental design prevents any confusion of regression effects in the model. However, this case is infrequent since the explanatory variables are not controlled (except in the case where the variables are set by the experimenter).

The presence of categorical predictors in the linear model requires the definition of new variables called: dummy variables. To handle the presence of categorical predictors and the redundancy of the model, two solutions can be considered: 1 - elimination of the intercept μ from the ANOVA model, 2- use of a contrast matrix (the constant

term is part of the model).

The most common contrast matrices are: treatment contrast (the most widely used by statisticians), sum contrast and Helmert contrast. The interpretation of the regression effects depends on the type of contrast considered. For example, the treatment contrast is recommended when comparing the level effects to a reference level.

What is the effect of an unbalanced experimental design on the quality of the estimation?

In the following part of this appendix, we will provide an answer for these questions on an illustrated example.

To achieve our purpose and for simplicity, we consider a one-way analysis variance of response Y$\in \mathbf{R}^n$ with factor "F" characterized by four levels of size n_i for i = 1, 2, 3, 4 respectively. A general expression of the model is:

$$Y = X\beta + E \quad E \sim N(0, \sigma^2)$$

Here, the design matrix X can be written in different ways:

- If we decide to remove the intercept, the design matrix X is written by using the binary coding X^*. In this case, dummy variables are created and the m-th observation of level j is associated $X^*[m,] = (0, \ldots, \mathbf{1}_j, \ldots, 0)$, where the only value of 1 corresponds to the jth column of matrix X*.
- If we decide to use a contrast matrix, the design matrix X is written as X = ($\mathbf{1} \quad X^*C$), where $\mathbf{1}$ is the unit vector and C is the contrast matrix. For example, the matrices corresponding to treatment C_t and sum C_s contrasts for a factor of 4 levels are written respectively as:

$$C_t = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad C_s = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix}$$

After constructing the design matrix X, there are many criteria that can be used to optimize the experimental design. Here, we choose to define n_1, n_2, n_3 and n_4 , the number of observations that should be collected for each level of factor F, by minimizing the sum of the variance for regression effects under the following constraint: $\sum_{i=1}^{4} n_i = n$.

The variance of the regression effect β is given by:

$$V(\beta) = (X'X)^{-1}\sigma^2$$

Thus, minimizing the sum of the variance of the regression effects consists in minimizing the trace of the matrix $(X'X)^{-1}$. Here, the Lagrange multiplier is used as a strategy to find the minima of the trace of the matrix $(X'X)^{-1}$. Thus, under the so-called **A** optimal design:

- For an ANOVA model without intercept, for a total number of observations n, the number of observations must be equiprobable between the 4 levels: $n_1 = n_2 = n_3 = n_4 = \frac{n}{4}$
- For an ANOVA model using contrasts:
  - For a treatment contrast, where the first level is taken as the default choice in R to be the baseline reference, the design is optimal if: $n_2 = n_3 = n_4 = \frac{n_1}{2}$, $\frac{2}{5}$ of the total number of observations n is collected for the baseline level and each of the remaining three levels should be given $\frac{1}{5}$ of n.

– For a sum contrast, the design is optimal if the number of observations collected for the baseline level is equal to $\frac{1}{1+3\sqrt{3}}$n and each of the remaining three levels should have $\frac{\sqrt{3}}{1+3\sqrt{3}}$n observations

How is the regression effect variance affected by the experimental design?

In order to illustrate the impact of an unbalanced experimental design on the quality of the estimation, we summarize in Tables 7.2, 7.3 and 7.4, the sum of variances of the regression effects under different experimental designs: balanced, unbalanced and strongly unbalanced for n = 20.

Experimental design	n_1	n_2	n_3	n_4	sum of regression effect variances
Balanced	5	5	5	5	0.80 σ^2
Unbalanced	4	3	6	7	0.89 σ^2
Strongly unbalanced	3	10	5	2	1.13 σ^2

Table 7.2: Sum of variances of the regression effects for the constant model under different experimental designs. n_1 refers to the number of observations for the baseline level. The total number of observations is set to 20.

Experimental design	n_1	n_2	n_3	n_4	sum of regression effect variances
Optimal	8	4	4	4	1.25 σ^2
Unbalanced	4	3	6	7	1.64 σ^2
Strongly unbalanced	3	10	5	2	2.13 σ^2

Table 7.3: Sum of variances of the regression effects under different experimental designs. A treatment contrast is used to design the matrix design. n_1 refers to the number of observations for the baseline level. The total number of observations is set to 20.

Experimental design	n_1	n_2	n_3	n_4	sum of regression effect variances
Optimal	3	5	6	6	0.48 σ^2
Unbalanced	4	3	6	7	0.54 σ^2
Strongly unbalanced	3	10	5	2	0.68 σ^2

Table 7.4: Sum of variances of the regression effects under different experimental designs. A sum contrast is used to design the matrix design. n_1 refers to the number of observations for the baseline level. The total number of observations is set to 20.

According to Tables 7.2, 7.3 and 7.4, an unbalanced experimental design with regards to the optimal one decreases the precision of the regression effects. For uncontrolled experiments (observational data), the design can be far away from the optimal design and the scientist may suffer from the poor precision of the regression effects.

7.4 References soil carbon database profiles

Agnelli, A., Trumbore, S. E., Corti, G., and Ugolini, F. (2002). The dynamics of organic matter in rock fragments in soil investigated by 14C dating and measurements of 13C. *European Journal of Soil Science*, 53(1):147–159.

Baisden, W., Amundson, R., Cook, A., and Brenner, D. (2002). Turnover and storage of C and N in five density fractions from California annual grassland surface soils. *Global Biogeochemical Cycles*, 16(4):64–1.

Balesdent, J. and Guillet, B. (1982). Les datations par le 14C des matières organiques des sols. contribution à l'étude de l'humification et du renouvellement des substances humiques. *Science du sol*, 2:93–111.

Basile-Doelsch, I., Amundson, R., Stone, W., Masiello, C., Bottero, J. Y., Colin, F., Masin, F., Borschneck, D., and Meunier, J. D. (2005). Mineralogical control of organic carbon dynamics in a volcanic ash soil on la Réunion. *European Journal of Soil Science*, 56(6):689–703.

Becker-Heidmann, P., Andresen, O., Kalmar, D., Scharpenseel, H.-W., and Yaalon, D. H. (2002). Carbon dynamics in vertisols as revealed by high-resolution sampling. *Radiocarbon*, 44(1):63–73.

Becker-Heidmann, P. and Scharpenseel, H.-W. (1986). Thin layer δ13C and D14C monitoring of "lessive" soil profiles. *Radiocarbon*, 28(2A):383–390.

Becker-Heidmann, P. and Scharpenseel, H. W. (1989). Carbon isotope dynamics in some tropical soils. *Radiocarbon*, 31(3):672–679.

Biedenbender, S. H., McClaran, M. P., Quade, J., and Weltz, M. A. (2004). Landscape patterns of vegetation change indicated by soil carbon isotope composition. *Geoderma*, 119(1-2):69–83.

Bol, R., Huang, Y., Meridith, J., Eglinton, G., Harkness, D., and Ineson, P. (1996). The 14C age and residence time of organic matter and its lipid constituents in a stagnohumic gley soil. *European Journal of Soil Science*, 47(2):215–222.

Butman, D., Raymond, P., Oh, N.-H., and Mull, K. (2007). Quantity, 14C age and lability of desorbed soil organic carbon in fresh water and seawater. *Organic Geochemistry*, 38(9):1547–1557.

Caner, L. and Bourgeon, G. (2001). Sur les possibilités de reconstitution paléo-environnementale offertes par les andosols des hautes terres tropicales. Exemple des Nilgiri (Inde du Sud). *Comptes Rendus de l'Académie des Sciences-Series IIA-Earth and Planetary Science*, 333(11):725–731.

Caner, L., Seen, D. L., Gunnell, Y., Ramesh, B. R., and Bourgeon, G. (2007). Spatial heterogeneity of land cover response to climatic change in the Nilgiri highlands (southern India) since the Last Glacial Maximum. *The Holocene*, 17(2):195–205.

Caner, L., Toutain, F., Bourgeon, G., and Herbillon, A.-J. (2003). Occurrence of sombric-like subsurface a horizons in some andic soils of the Nilgiri Hills (Southern India) and their palaeoecological significance. *Geoderma*, 117(3-4):251–265.

Chabbi, A., Kögel-Knabner, I., and Rumpel, C. (2009). Stabilised carbon in subsoil horizons is located in spatially distinct parts of the soil profile. *Soil Biology and Biochemistry*, 41(2):256–261.

Chen, Q., Sun, Y., Shen, C., Peng, S., Yi, W., Li, Z., and Jiang, M. (2002). Organic matter turnover rates and CO2 flux from organic matter decomposition of mountain soil profiles in the subtropical area, south China. *Catena*, 49(3):217–229.

Cherkinsky, A. (1996). 14C Dating and Soil Organic Matter Dynamics in Arctic and Subarctic Ecosystems. *Radiocarbon*, 38(2):241–245.

Cherkinsky, A. and Brovkin, V. (1993). Dynamics of radiocarbon in soils. *Radiocarbon*, 35(3):363–367.

Chiti, T., Neubert, R., Janssens, I., Certini, G., Yuste, J. C., and Sirignano, C. (2009). Radiocarbon dating reveals different past managements of adjacent forest soils in the Campine region, Belgium. *Geoderma*, 149(1-2):137–142.

Chiti, T., Papale, D., Smith, P., Dalmonech, D., Matteucci, G., Yeluripati, J., Rodeghiero, M., and Valentini, R. (2010). Predicting changes in soil organic carbon in mediterranean and alpine forests during the Kyoto Protocol commitment periods using the CENTURY model. *Soil use and management*, 26(4):475–484.

Conen, F., Karhu, K., Leifeld, J., Seth, B., Vanhala, P., Liski, J., and Alewell, C. (2008). Temperature sensitivity of young and old soil carbon–same soil, slight differences in 13C natural abundance method, inconsistent results. *Soil Biology and Biochemistry*, 40(10):2703–2705.

de Freitas, H. A., Pessenda, L. C. R., Aravena, R., Gouveia, S. E. M., de Souza Ribeiro, A., and Boulet, R. (2001). Late quaternary vegetation dynamics in the southern Amazon basin inferred from carbon isotopes in soil organic matter. *Quaternary Research*, 55(1):39–46.

Desjardins, T., Andreux, F., Volkoff, B., and Cerri, C. (1994). Organic carbon and 13C contents in soils and soil size-fractions, and their changes due to deforestation and pasture installation in eastern Amazonia. *Geoderma*, 61(1-2):103–118.

Dörr, H. and Münnich, K. (1989). Downward movement of soil organic matter and its influence on trace-element transport (210 Pb, 137 Cs) in the soil. *Radiocarbon*, 31(3):655–663.

Dümig, A., Schad, P., Rumpel, C., Dignac, M.-F., and Kögel-Knabner, I. (2008). Araucaria forest expansion on grassland in the southern Brazilian highlands as revealed by 14C and δ13C studies. *Geoderma*, 145(1-2):143–157.

Elzein, A. and Balesdent, J. (1995). Mechanistic simulation of vertical distribution of carbon concentrations and residence times in soils. *Soil Science Society of America Journal*, 59(5):1328–1335.

Favilli, F., Egli, M., Cherubini, P., Sartori, G., Haeberli, W., and Delbos, E. (2008). Comparison of different methods of obtaining a resilient organic matter fraction in Alpine soils. *Geoderma*, 145(3-4):355–369.

Gillson, L., Waldron, S., and Willis, K. (2004). Interpretation of soil δ13C as an indicator of vegetation change in African savannas. *Journal of Vegetation Science*, 15(3):339–350.

Guillet, B., Disnar, J.-R., Loustau, D., and Balesdent, J. (2010). Dynamics of soil carbon and moder horizons related to age in pine and beech stands. *Forests, Carbon Cycle and Climate Change*, 55.

Harkness, D., Harrison, A., and Bacon, P. (1986). The temporal distribution of 'bomb'14C in a forest soil. *Radiocarbon*, 28(2A):328–337.

Huang, Y., Li, B., Bryant, C., Bol, R., and Eglinton, G. (1999). Radiocarbon dating of aliphatic hydrocarbons a new approach for dating passive-fraction carbon in soil horizons. *Soil Science Society of America Journal*, 63(5):1181–1187.

Huang, Y., Roland Bol, D. D. H., and Philip Ineson, G. E. (1996). Post-glacial variations in distributions, 13C and 14C contents of aliphatic hydrocarbons and bulk organic matter in three types of British acid upland soils. *Organic Geochemistry*, 24(3):273–287.

Jagercikova, M., Cornu, S., Bourlès, D., Evrard, O., Hatté, C., and Balesdent, J. (2017). Quantification of vertical solid matter transfers in soils during pedogenesis by a multi-tracer approach. *Journal of soils and sediments*, 17(2):408–422.

Katsuno, K., Miyairi, Y., Tamura, K., Matsuzaki, H., and Fukuda, K. (2010). A study of the carbon dynamics of Japanese grassland and forest using 14C and 13C. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 268(7-8):1106–1109.

Koarashi, J., Iida, T., and Asano, T. (2005). Radiocarbon and stable carbon isotope compositions of chemically fractionated soil organic matter in a temperate-zone forest. *Journal of environmental radioactivity*, 79(2):137–156.

Kondo, M., Uchida, M., and Shibata, Y. (2010). Radiocarbon-based residence time estimates of soil organic carbon in a temperate forest: Case study for the density fractionation for japanese volcanic ash soil. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 268(7-8):1073–1076.

Kovda, I., Lynn, W., Williams, D., and Chichagova, O. (2001). Radiocarbon age of Vertisols and its interpretation using data on gilgai complex in the north Caucasus. *Radiocarbon*, 43(2B):603–610.

Krull, E. S., Bestland, E. A., Skjemstad, J. O., and Parr, J. F. (2006). Geochemistry (δ13C, δ15N, 13C NMR) and residence times (14C and OSL) of soil organic matter from red-brown earths of South Australia: Implications for soil genesis. *Geoderma*, 132(3-4):344–360.

Krull, E. S. and Skjemstad, J. O. (2003). δ13C and δ15N profiles in 14C-dated Oxisol and Vertisols as a function of soil chemistry and mineralogy. *Geoderma*, 112(1-2):1–29.

Krull, E. S., Skjemstad, J. O., Burrows, W. H., Bray, S. G., Wynn, J. G., Bol, R., Spouncer, L., and Harms, B. (2005). Recent vegetation changes in central Queensland, Australia: Evidence from δ13C and 14C analyses of soil organic matter. *Geoderma*, 126(3-4):241–259.

Ladyman, S. and Harkness, D. (1980). Carbon isotope measurement as an index of soil development. *Radiocarbon*, 22(3):885–891.

Laskar, A. H., Yadava, M., and Ramesh, R. (2012). Radiocarbon and stable carbon isotopes in two soil profiles from northeast India. *Radiocarbon*, 54(1):81–89.

Lassey, K., Tate, K., Sparks, R., and Claydon, J. (1996). Historic measurements of radiocarbon in New Zealand soils. *Radiocarbon*, 38(2):253–270.

Leavitt, S., Follett, R., Kimble, J., and Pruessner, E. (2007). Radiocarbon and δ13C depth profiles of soil organic carbon in the US great plains: A possible spatial record of paleoenvironment and paleovegetation. *Quaternary International*, 162:21–34.

Liu, W., Moriizumi, J., Yamazawa, H., and Iida, T. (2006). Depth profiles of radiocarbon and carbon isotopic compositions of organic matter and CO2 in a forest soil. *Journal of environmental radioactivity*, 90(3):210–223.

Lobo, P., Flexor, J., Rapaire, J., and Sieffermann, G. (1974). Determination of the time of residence of humic fractions of two ferallitic soils using natural and thermonuclear radiocarbon. *Cah. ORSTOM, Ser. Pedol*, 12(1):115–123.

Mariotti, A. and Peterschmitt, E. (1994). Forest savanna ecotone dynamics in India as revealed by carbon isotope ratios of soil organic matter. *Oecologia*, 97(4):475–480.

Martel, Y. and Paul, E. A. (1974). Effects of cultivation on the organic matter of grassland soils as determined by fractionation and radiocarbon dating. *Canadian Journal of Soil Science*, 54(4):419–426.

Mayer, J. H., Burr, G. S., and Holliday, V. T. (2008). Comparisons and interpretations of charcoal and organic matter radiocarbon ages from buried soils in north-central Colorado, USA. *Radiocarbon*, 50(3):331–346.

McClaran, M. P. and Umlauf, M. (2000). Desert grassland dynamics estimated from carbon isotopes in grass phytoliths and soil organic matter. *Journal of Vegetation Science*, 11(1):71–76.

McClung de Tapia, E. (2005). Radiocarbon dates from soil profiles in the Teotihuacan Valley, Mexico: indicators of geomorphological processes. *Radiocarbon*, 47:159–175.

Milton, G. and Kramer, S. (1997). Using 14C as a tracer of carbon accumulation and turnover in soils. *Radiocarbon*, 40(2):999–1011.

O'Brien, B. J. (1984). Soil organic carbon fluxes and turnover rates estimated from radiocarbon enrichment. *Soil Biology and Biochemistry*, 16(2):115–120.

O'Brien, B. J. (1986). The use of natural and anthropogenic 14C to investigate the dynamics of soil organic carbon. *Radiocarbon*, 28(2A):358–362.

Paul, E., Collins, H., and Leavitt, S. (2001). Dynamics of resistant soil carbon of Midwestern agricultural soils measured by naturally occurring 14C abundance. *Geoderma*, 104(3-4):239–256.

Paul, E., Follett, R., Leavitt, S., Halvorson, A., Peterson, G., and Lyon, D. (1997). Radiocarbon dating for determination of soil organic matter pool sizes and dynamics. *Soil Science Society of America Journal*, 61(4):1058–1067.

Pessenda, L., Gomes, B., Aravena, R., Ribeiro, A., Boulet, R., and Gouveia, S. (1998). The carbon isotope record in soils along a forest-cerrado ecosystem transect: implications for vegetation changes in the Rondonia state, southwestern Brazilian Amazon region. *The Holocene*, 8(5):599–603.

Pessenda, L., Gouveia, S., and Aravena, R. (2001). Radiocarbon dating of total soil organic matter and humin fraction and its comparison with 14C ages of fossil charcoal. *Radiocarbon*, 43(2B):595–601.

Pessenda, L. C., Aravena, R., Melfi, A. J., Telles, E. C., Boulet, R., Valencia, E., and Tomazello, M. (1996). The use of carbon isotopes (13C, 14C) in soil to evaluate vegetation changes during the Holocene in central Brazil. *Radiocarbon*, 38(2):191–201.

Rasmussen, C., Torn, M. S., and Southard, R. J. (2005). Mineral assemblage and aggregates control carbon dynamics in a California conifer forest. *Soil Science Society of America Journal*, 69(6):1711–1721.

Rethemeyer, J., Kramer, C., Gleixner, G., John, B., Yamashita, T., Flessa, H., Andersen, N., Nadeau, M.-J., and Grootes, P. M. (2005). Radiocarbon analysis of functional-defined and molecular organic matter fractions from agricultural soil profiles. *Geoderma*, 128:94–105.

Rethemeyer, J., Kramer, C., Gleixner, G., Wiesenberg, G. L., Schwark, L., Andersen, N., Nadeau, M.-J., and Grootes, P. M. (2004). Complexity of soil organic matter: AMS 14C analysis of soil lipid fractions and individual compounds. *Radiocarbon*, 46(1):465–473.

Rumpel, C., Chaplot, V., Chabbi, A., Largeau, C., and Valentin, C. (2008). Stabilisation of HF soluble and HCl resistant organic matter in sloping tropical soils under slash and burn agriculture. *Geoderma*, 145(3-4):347–354.

Rumpel, C., Kögel-Knabner, I., and Bruhn, F. (2002). Vertical distribution, age, and chemical composition of organic carbon in two forest soils of different pedogenesis. *Organic Geochemistry*, 33(10):1131–1142.

Scharpenseel, H. and Pietig, F. (1973a). University of Bonn natural radiocarbon measurements V. *Radiocarbon*, 15(1):13–41.

Scharpenseel, H. W. and Pietig, F. (1973b). University of bonn natural radiocarbon measurements vi. *Radiocarbon*, 15(2):252–279.

Schulze, K., Borken, W., Muhr, J., and Matzner, E. (2009). Stock, turnover time and accumulation of organic matter in bulk and density fractions of a podzol soil. *European Journal of Soil Science*, 60(4):567–577.

Schwartz, D., Mariotti, A., Trouvé, C., Van Den Borg, K., and Guillet, B. (1992). Etude des profils isotopiques 13C et 14C d'un sol ferrallitique sableux du littoral congolais: implications sur la dynamique de la matière organique et l'histoire de la végétation. *Comptes Rendus de l'Académie des Sciences. Série 2: Mécanique...*, 315:1411–1417.

Shen, C., Yi, W., Sun, Y., Xing, C., Yang, Y., Yuan, C., Li, Z., Peng, S., An, Z., and Liu, T. (2001). Distribution of 14C and 13C in forest soils of the Dinghushan Biosphere Reserve. *Radiocarbon*, 43(2B):671–678.

Stephan, S., Berrier, J., De Petre, A., Jeanson, C., Kooistra, M., Scharpenseel, H., and Schiffmann, H. (1983). Characterization of in situ organic matter constituents in vertisols from Argentina, using submicroscopic and cytochemical methods—first report. *Geoderma*, 30(1-4):21–34.

Stout, J. and Goh, K. (1980). The use of radiocarbon to measure the effects of earthworms on soil development. *Radiocarbon*, 22(3):892–896.

Tefs, C. and Gleixner, G. (2012). Importance of root derived carbon for soil organic matter storage in a temperate old-growth beech forest–evidence from C, N and 14C content. *Forest Ecology and Management*, 263:131–137.

Tegen, I. and Dörr, H. (1996). 14C measurements of soil organic matter, soil CO2 and dissolved organic carbon (1987–1992). *Radiocarbon*, 38(2):247–251.

Tonneijck, F. H., van der Plicht, J., Jansen, B., Verstraten, J. M., and Hooghiemstra, H. (2006). Radiocarbon dating of soil organic matter fractions in Andosols in northern Ecuador. *Radiocarbon*, 48(3):337–353.

Torn, M. S., Lapenis, A. G., Timofeev, A., Fischer, M. L., Babikov, B. V., and Harden, J. W. (2002). Organic carbon and carbon isotopes in modern and 100-year-old-soil archives of the Russian steppe. *Global Change Biology*, 8(10):941–953.

Van Dam, Douwe, N. v. B. and Veldkamp, E. (1997). Soil organic carbon dynamics: variability with depth in forested and deforested soils under pasture in Costa Rica. *Biogeochemistry*, 39(3):343–375.

Van Mourik, J., Nierop, K. G., and Vandenberghe, D. (2010). Radiocarbon and optically stimulated luminescence dating based chronology of a polycyclic driftsand sequence at Weerterbergen (SE Netherlands). *Catena*, 80(3):170–181.

Wang, G., Han, J., Zhou, L., Xiong, X., and Wu, Z. (2005). Carbon isotope ratios of plants and occurrences of C4 species under different soil moisture regimes in arid region of Northwest China. *Physiologia Plantarum*, 125(1):74–81.

Wang, Yang, R. A. and Trumbore, S. (1999). The impact of land use change on C turnover in soils. *Global Biogeochemical Cycles*, 13(1):47–57.

Wang, Y., Amundson, R., and Trumbore, S. (1996). Radiocarbon dating of soil organic matter. *Quaternary Research*, 45(3):282–288.

site_nb	MAT	MAP	min_P	AI	Dif_T	F14Catm	WRB	Ecos	country	long	lat	Ref_paper
1	21.90	1733.20	27.30	1.50	15.20	113.40	Fluvisol	natural-forest	China	112.51	23.16	(Shen et al., 2001)
2	21.50	1744.70	28.90	1.54	15.50	113.40	Fluvisol	natural-forest	China	112.53	23.17	(Shen et al., 2001)
3	21.50	1744.70	28.90	1.54	15.50	113.40	Fluvisol	natural-forest	China	112.56	23.19	(Shen et al., 2001)
7	21.50	1744.70	28.90	1.54	15.50	97.20	Fluvisol	natural-forest	China	112.53	23.17	(Chen et al., 2002)
8	21.50	1744.70	28.90	1.54	15.50	97.20	Fluvisol	natural-grassland	China	112.53	23.17	(Chen et al., 2002)
19	6.40	722.50	45.20	1.15	18.70	105.60	Plinthosol	natural-forest	Germany	11.85	50.10	(Rumpel et al., 2002)
20	8.40	673.30	41.90	1.04	18.50	105.60	Podzol	natural-forest	Germany	10.46	49.87	(Rumpel et al., 2002)
45	2.80	858.10	4.00	1.00	16.00	68.90	Podzol	natural-forest	China	102.00	29.50	Wang et al. (2005)
46	2.80	858.10	4.00	1.00	16.00	68.90	Podzol	natural-forest	China	102.00	29.50	Wang et al. (2005)
47	2.80	858.10	4.00	1.00	16.00	68.90	Plinthosol	natural-grassland	China	102.00	29.50	Wang et al. (2005)
48	2.80	858.10	4.00	1.00	16.00	68.90	Podzol	natural-grassland	China	102.00	29.50	Wang et al. (2005)
63	20.50	1269.90	25.40	1.00	6.20	121.90	Fluvisol	natural-forest	Brazil	-47.63	-22.72	(Pessenda et al., 1996)
64	26.30	2039.60	42.70	1.59	1.40	146.70	Leptosol	natural-forest	Brazil	-52.97	-3.50	(Pessenda et al., 1996)
67	24.60	2004.10	9.90	1.45	4.10	146.70	Fluvisol	undefined-natural	Brazil	-60.12	-12.70	(Pessenda et al., 1998)
68	26.20	1985.80	10.80	1.43	3.80	146.70	Fluvisol	natural-forest	Brazil	-61.17	-11.82	(Pessenda et al., 1998)
69	26.20	1985.80	10.80	1.43	3.80	146.70	Fluvisol	natural-forest	Brazil	-61.25	-11.77	(Pessenda et al., 1998)
70	26.00	2040.10	14.40	1.55	3.00	146.70	Leptosol	natural-forest	Brazil	-62.82	-10.17	(Pessenda et al., 1998)
71	26.40	2237.40	25.00	1.75	1.40	146.70	PALEOSOL	undefined-natural	Brazil	-63.03	-7.52	(Pessenda et al., 1998)
73	26.40	2237.40	25.00	1.75	1.40	146.70	Podzol	natural-forest	Brazil	-63.03	-7.52	(Pessenda et al., 1998)
74	26.40	2237.40	25.00	1.75	1.40	146.70	Podzol	natural-forest	Brazil	-63.03	-7.52	(Pessenda et al., 1998)
75	26.40	2237.40	25.00	1.75	1.40	146.70	Leptosol	natural-forest	Brazil	-63.03	-7.52	(Pessenda et al., 1998)
82	14.40	648.40	18.00	0.49	10.80	88.90	Leptosol	field_cultivated	Australia	140.80	-37.30	(Krull et al., 2006)
83	5.50	1664.90	92.00	4.22	11.60	150.60	Anthrosol	undefined-natural	UK	-2.45	54.68	(Huang et al., 1996)
84	5.50	1664.90	92.00	4.22	11.60	150.60	Podzol	undefined-natural	UK	-2.45	54.68	(Huang et al., 1996)
85	5.50	1664.90	92.00	4.22	11.60	150.60	Plinthosol	undefined-natural	UK	-2.45	54.68	(Huang et al., 1996)
91	13.80	475.80	12.30	0.28	23.70	113.40	Andosol	natural-grassland	USA	-102.08	35.17	(Leavitt et al., 2007)
92	9.40	382.10	7.60	0.28	26.80	118.70	Andosol	natural-grassland	USA	-103.29	39.94	(Leavitt et al., 2007)
93	12.70	436.00	8.70	0.26	24.40	113.40	Histosol	natural-grassland	USA	-102.54	36.08	(Leavitt et al., 2007)
94	4.20	518.20	10.80	0.57	36.30	105.60	Andosol	natural-grassland	USA	-96.45	47.93	(Leavitt et al., 2007)
95	7.20	724.60	18.10	0.69	34.20	105.60	Andosol	natural-grassland	USA	-94.30	44.72	(Leavitt et al., 2007)
96	9.70	872.70	24.50	0.78	30.70	118.70	Andosol	natural-grassland	USA	-93.66	41.20	(Leavitt et al., 2007)
97	5.40	407.70	8.70	0.40	34.00	106.40	Andosol	natural-grassland	USA	-101.00	46.74	(Leavitt et al., 2007)
98	4.50	452.00	10.50	0.47	35.80	106.40	Andosol	natural-grassland	USA	-99.28	46.95	(Leavitt et al., 2007)
99	12.00	988.50	38.40	0.85	28.30	113.40	Histosol	natural-grassland	USA	-92.00	38.95	(Leavitt et al., 2007)
100	5.80	342.40	7.30	0.31	33.70	113.40	Andosol	natural-grassland	USA	-104.68	47.70	(Leavitt et al., 2007)
101	10.30	700.80	14.50	0.59	30.90	118.70	Andosol	natural-grassland	USA	-97.06	40.89	(Leavitt et al., 2007)
102	3.00	505.60	10.90	0.59	36.50	105.60	Andosol	natural-grassland	USA	-96.22	48.64	(Leavitt et al., 2007)
107	19.70	681.60	28.30	0.41	14.20	106.40	Acrisol	grassland_cultivated	Australia	150.60	-26.72	(Krull and Skjemstad, 2003)
108	19.50	665.50	31.90	0.40	14.30	106.40	Acrisol	grassland_cultivated	Australia	150.52	-27.45	(Krull and Skjemstad, 2003)

site_nb	MAT	MAP	min_P	AI	Dif_T	F14Catm	WRB	Ecos	country	long	lat	Ref_paper
109	18.00	1655.10	49.10	1.06	10.00	106.40	Fluvisol	natural-forest	Australia	153.20	-28.17	(Krull and Skjemstad, 2003)
110	0.40	1008.10	53.30	1.80	14.70	74.50	Plinthosol	natural-forest	Italy	10.75	46.35	(Favilli et al., 2008)
111	0.40	1008.10	53.30	1.80	14.70	74.50	Plinthosol	natural-forest	Italy	10.75	46.40	(Favilli et al., 2008)
114	22.50	1641.40	13.80	1.35	7.90	68.90	Podzol	field_cultivated	Laos	102.35	19.85	(Rumpel et al., 2008)
115	22.50	1641.40	13.80	1.35	7.90	68.90	Histosol	field_cultivated	Laos	102.35	19.85	(Rumpel et al., 2008)
116	22.50	1641.40	13.80	1.35	7.90	68.90	Podzol	natural-forest	Laos	102.35	19.85	(Rumpel et al., 2008)
130	7.60	826.90	48.40	1.02	26.70	134.20	Plinthosol	grassland_cultivated	Canada	-79.07	43.81	(Milton and Kramer, 1997)
131	4.20	820.80	48.20	1.13	31.90	134.20	Plinthosol	forest_cultivated	Canada	-77.40	46.00	(Milton and Kramer, 1997)
132	4.40	938.20	54.00	1.29	31.70	134.20	Anthrosol	forest_cultivated	Canada	-79.47	46.31	Milton_1998_Radiocarbon
133	26.50	890.90	3.80	0.50	11.20	226.40	Histosol	field_cultivated	India	78.27	17.53	(Becker-Heidmann and Scharpenseel, 1989)
134	26.30	2234.20	33.70	1.59	3.20	226.40	Andosol	field_cultivated	Philippines	121.32	13.95	(Becker-Heidmann and Scharpenseel, 1989)
135	27.30	2386.00	42.60	1.61	3.60	226.40	Anthrosol	field_cultivated	Philippines	121.22	14.17	(Becker-Heidmann and Scharpenseel, 1989)
136	27.20	2045.50	55.00	1.35	2.70	226.40	Acrisol	field_cultivated	Philippines	123.08	11.35	(Becker-Heidmann and Scharpenseel, 1989)
137	19.60	517.50	0.00	0.36	13.80	189.70	Acrisol	field_cultivated	Israel	34.78	31.68	(Becker-Heidmann et al., 2002)
138	18.80	694.50	0.00	0.50	15.20	200.20	Acrisol	grassland_cultivated	Israel	35.25	32.90	(Becker-Heidmann et al., 2002)
139	26.50	890.90	3.80	0.50	11.20	226.40	Acrisol	field_cultivated	India	78.28	17.58	(Becker-Heidmann et al., 2002)
143	10.00	792.40	49.40	1.14	15.10	68.90	Luvisol	forest_cultivated	Belgium	4.52	51.31	(Chiti et al., 2009)
146	3.60	1387.70	87.90	2.29	16.20	59.20	Arenosol	natural-grassland	Switzerland	7.92	46.25	(Conen et al., 2008)
147	3.60	1387.70	87.90	2.29	16.20	59.20	Arenosol	natural-grassland	Switzerland	7.92	46.25	(Conen et al., 2008)
148	3.60	1387.70	87.90	2.29	16.20	59.20	Arenosol	natural-grassland	Switzerland	7.92	46.25	(Conen et al., 2008)
164	5.50	1664.90	92.00	4.22	11.60	125.70	Plinthosol	undefined-natural	UK	-2.45	54.68	(Huang et al., 1999)
165	5.50	1664.90	92.00	4.22	11.60	125.70	Podzol	undefined-natural	UK	-2.45	54.68	(Huang et al., 1999)
166	5.50	1664.90	92.00	4.22	11.60	125.70	Anthrosol	undefined-natural	UK	-2.45	54.68	(Huang et al., 1999)
167	5.50	1664.90	92.00	4.22	11.60	125.70	Anthrosol	undefined-natural	UK	-2.45	54.68	(Bol et al., 1996)
168	10.90	1112.90	44.60	1.32	26.50	80.10	Regosol	natural-forest	japan	138.21	36.52	(Katsuno et al., 2010)
169	5.90	1506.30	41.00	2.12	25.50	80.10	Regosol	grassland_cultivated	Japan	138.35	36.52	(Katsuno et al., 2010)
170	15.90	1639.50	37.90	1.59	23.40	80.10	Regosol	natural-forest	Japan	136.97	35.15	(Koarashi et al., 2005)
175	25.00	1213.10	0.00	1.00	5.10	194.50	Fluvisol	natural-grassland	Congo	11.95	-4.72	(Schwartz et al., 1992)
190	10.20	1109.80	78.00	2.04	24.60	68.90	Plinthosol	natural-forest	USA	-72.75	41.26	(Butman et al., 2007)
193	6.40	722.50	45.20	1.15	18.70	56.20	Plinthosol	forest_cultivated	Germany	11.87	50.13	(Schulze et al., 2009)
194	6.40	722.50	45.20	1.15	18.70	51.30	Plinthosol	forest_cultivated	Germany	11.87	50.13	(Schulze et al., 2009)
195	6.40	722.50	45.20	1.15	18.70	51.30	Plinthosol	forest_cultivated	Germany	11.87	50.13	(Schulze et al., 2009)
196	8.50	749.10	43.80	1.24	16.40	212.30	Histosol	natural-forest	Germany	10.15	53.72	(Becker-Heidmann and Scharpenseel, 1986)
197	8.90	706.70	43.80	1.13	16.80	212.30	Histosol	natural-forest	Germany	9.70	52.28	(Becker-Heidmann and Scharpenseel, 1986)
202	12.90	960.70	47.90	1.14	14.10	68.90	Plinthosol	natural-forest	France	-0.77	44.70	(Guillet et al., 2010)
203	12.90	973.20	48.50	1.14	14.20	68.90	Plinthosol	natural-forest	France	-0.58	44.64	(Guillet et al., 2010)
204	10.40	633.70	45.90	0.89	14.90	106.40	Histosol	natural-forest	France	1.17	48.33	(Guillet et al., 2010)
205	10.40	633.70	45.90	0.89	14.90	106.40	Histosol	natural-forest	France	1.17	48.33	(Guillet et al., 2010)r
206	10.40	633.70	45.90	0.89	14.90	74.50	Histosol	natural-forest	France	1.17	48.33	(Guillet et al., 2010)

site_nb	MAT	MAP	min_P	AI	Dif_T	F14Catm	WRB	Ecos	country	long	lat	Ref_paper
207	10.40	633.70	45.90	0.89	14.90	47.80	Histosol	natural-forest	France	1.17	48.33	(Guillet et al., 2010)
211	10.80	636.20	44.00	0.91	14.90	125.70	Histosol	natural-forest	France	2.05	48.87	(Elzein and Balesdent, 1995)
212	10.80	636.20	44.00	0.91	14.90	118.70	Podzol	natural-forest	France	2.05	48.87	(Balesdent 1994 unpublished)
214	6.80	1120.90	82.50	1.59	17.40	254.60	Histosol	grassland_cultivated	France	5.95	46.28	(Balesdent and Guillet, 1982)
215	6.80	1120.90	82.50	1.59	17.40	265.20	Andosol	grassland_cultivated	France	5.83	46.30	(Balesdent and Guillet, 1982)
216	6.80	1120.90	82.50	1.59	17.40	265.20	Histosol	grassland_cultivated	France	5.83	46.30	(Balesdent and Guillet, 1982)
217	24.90	4285.30	0.20	2.79	3.90	150.60	Fluvisol	natural-forest	India	74.74	13.95	Mariotti_1994_Oecologia
221	5.50	507.30	28.50	0.62	31.20	106.40	Andosol	natural-grassland	Russia	43.56	51.73	Torn et al. (2002)
222	5.50	507.30	28.50	0.62	31.20	105.60	Andosol	natural-grassland	Russia	43.56	51.73	Torn et al. (2002)
223	10.50	1716.60	31.00	1.98	24.10	65.20	Regosol	natural-forest	Japan	137.57	35.22	(Liu et al., 2006)
224	10.50	1716.60	31.00	1.98	24.10	65.20	Regosol	natural-forest	Japan	137.57	35.22	(Liu et al., 2006)
225	10.50	1716.60	31.00	1.98	24.10	65.20	Regosol	natural-forest	Japan	137.57	35.22	(Liu et al., 2006)
226	10.50	1716.60	31.00	1.98	24.10	65.20	Regosol	natural-forest	Japan	137.57	35.22	(Liu et al., 2006)
227	17.00	1708.80	12.40	1.35	3.80	113.40	Regosol	natural-forest	India	76.60	11.30	(Caner and Bourgeon, 2001)
228	16.10	1552.30	11.10	1.24	3.90	113.40	Regosol	forest_cultivated	India	76.58	11.42	(Caner and Bourgeon, 2001)
230	7.10	824.20	27.50	0.89	30.90	113.40	Andosol	field_cultivated	United State	-89.35	43.30	(Paul et al., 2001)
231	8.10	765.10	34.30	0.85	27.90	113.40	Podzol	field_cultivated	United State	-84.12	43.38	(Paul et al., 2001)
232	9.50	904.10	40.40	0.91	27.70	113.40	Histosol	natural-forest	United State	-85.50	42.30	(Paul et al., 2001)
233	9.60	901.20	46.70	0.89	27.70	113.40	Histosol	field_cultivated	United State	-84.00	41.00	(Paul et al., 2001)
234	10.40	982.20	58.30	0.95	26.40	113.40	Histosol	field_cultivated	United State	-83.50	39.80	(Paul et al., 2001)
235	10.20	1127.70	70.60	1.10	26.00	113.40	Histosol	field_cultivated	United State	-75.72	40.55	(Paul et al., 2001)
237	17.00	1708.80	12.40	1.35	3.80	105.60	Regosol	forest_cultivated	India	76.62	11.22	(Caner et al., 2003)
238	16.70	1429.80	11.50	1.13	4.00	105.60	Regosol	forest_cultivated	India	76.82	11.42	(Caner et al., 2003)
240	1.30	396.20	14.20	0.51	35.80	574.10	Andosol	field_cultivated	Canada	-106.60	52.60	(Martel and Paul, 1974)
241	1.30	396.20	14.20	0.51	35.80	574.10	Anthrosol	field_cultivated	Canada	-106.60	52.60	(Martel and Paul, 1974)
243	1.80	390.50	12.60	0.49	35.30	574.10	Anthrosol	undefined-natural	Canada	-104.40	51.40	(Martel and Paul, 1974)
244	9.60	395.70	7.40	0.29	26.90	134.20	Histosol	natural-grassland	United States	-103.20	40.20	(Paul et al., 1997)
247	9.60	395.70	7.40	0.29	26.90	134.20	Gleyso	field_cultivated	United States	-103.20	40.20	(Paul et al., 1997)
256	24.20	1562.90	22.70	1.22	2.40	134.20	Fluvisol	undefined-natural	Cameroon	13.73	4.33	(Guillet et al., 2010)
257	24.20	1562.90	22.70	1.22	2.40	113.40	Fluvisol	natural-forest	Cameroon	13.73	4.33	(Guillet et al., 2010)
258	24.10	3047.90	9.10	2.60	10.20	45.10	UNKNOW	field_cultivated	India	92.57	24.68	(Laskar et al., 2012)
259	24.10	3047.90	9.10	2.60	10.20	45.10	Plinthosol	natural-grassland	India	92.57	24.68	(Laskar et al., 2012)
260	15.20	432.80	4.80	0.26	18.10	125.70	Regosol-Arenosol	undefined-natural	USA	-110.63	31.75	(McClung de Tapia, 2005)
261	26.40	2467.30	43.40	1.78	1.70	171.00	Plinthosol	natural-forest	Brasil	-47.15	-1.73	(Desjardins et al., 1994)
270	11.40	582.40	36.90	0.66	24.70	534.20	calcisol	natural-forest	Bulgaria	24.58	43.37	(Scharpenseel and Pietig, 1973b)
271	11.60	585.20	36.40	0.66	24.60	534.20	calcisol	natural-forest	Bulgaria	24.29	43.53	(Scharpenseel and Pietig, 1973b)
272	12.00	591.00	33.90	0.64	21.60	534.20	Nitisol	natural-forest	Bulgaria	26.97	42.60	(Scharpenseel and Pietig, 1973b)
275	12.10	636.70	28.10	0.67	21.10	534.20	Histosol	natural-forest	Bulgaria	25.83	41.78	(Scharpenseel and Pietig, 1973b)
305	8.00	935.30	46.20	1.39	19.00	474.90	Acrisol	natural-forest	Germany	11.49	47.78	(Scharpenseel and Pietig, 1973a)

site_nb	MAT	MAP	min_P	AI	Dif_T	F14Catm	WRB	Ecos	country	long	lat	Ref_paper
306	9.90	698.30	48.10	1.05	14.80	45.10	Histosol	natural-grassland	France	3.03	49.87	(Jagercikova et al., 2017)
309	10.50	639.30	44.30	0.94	14.80	47.80	Histosol	natural-forest	France	1.97	48.90	J(Jagercikova et al., 2017)
330	25.30	1271.50	0.00	1.06	13.70	43.20	Luvisol	undefined-natural	Congo	11.75	-4.35	(Christine Hatté unpublished)
331	25.30	1271.50	0.00	1.06	5.00	51.00	Luvisol	forest_cultivated	Congo	11.75	-4.35	(Christine Hatté unpublished)
332	15.50	1753.20	69.00	1.65	6.00	94.90	Regosol	natural-forest	France	55.36	-21.08	(Basile-Doelsch et al., 2005)
333	16.00	428.00	0.80	0.27	18.90	106.40	Podzol	natural-grassland	USA	-120.46	37.51	(Baisden et al., 2002)
334	16.00	428.00	0.80	0.27	18.90	320.40	Podzol	natural-grassland	USA	-120.46	37.51	(Baisden et al., 2002)
335	16.00	428.00	0.80	0.27	18.90	-22.00	Podzol	natural-grassland	USA	-120.46	37.51	(Baisden et al., 2002)
336	16.00	369.20	0.80	0.23	18.40	106.40	Histosol	natural-grassland	USA	-120.59	37.52	(Baisden et al., 2002)
337	16.00	369.20	0.80	0.23	18.40	-19.90	Histosol	natural-grassland	USA	-120.59	37.52	(Baisden et al., 2002)
338	16.00	369.20	0.80	0.23	18.40	320.40	Histosol	natural-grassland	USA	-120.59	37.52	(Baisden et al., 2002)
339	16.00	369.20	0.80	0.23	18.40	106.40	Histosol	natural-grassland	USA	-120.59	37.63	(Baisden et al., 2002)
343	16.00	371.60	0.80	0.24	18.90	320.40	Histosol	natural-grassland	USA	-120.37	37.46	(Baisden et al., 2002)
344	16.00	371.60	0.80	0.24	18.90	106.40	Histosol	natural-grassland	USA	-120.37	37.46	(Baisden et al., 2002)

BIBLIOGRAPHY

Alexander, E. (1980). Bulk densities of california soils in relation to other soil properties. Soil Science Society of America Journal, 44(4):689–692.

Ali, S., Begum, F., Hayat, R., and Bohannan, B. J. (2017). Variation in soil organic carbon stock in different land uses and altitudes in bagrot valley, northern karakoram. Acta Agriculturae Scandinavica, Section B—Soil & Plant Science, 67(6):551–561.

Anderson, J. (1973). Carbon dioxide evolution from two temperate, deciduous woodland soils. Journal of Applied Ecology, pages 361–378.

Atchadé, Y., Fort, G., et al. (2010). Limit theorems for some adaptive mcmc algorithms with subgeometric kernels. Bernoulli, 16(1):116–154.

Balesdent, J., Basile-Doelsch, I., Chadoeuf, J., Cornu, S., Derrien, D., Fekiacova, Z., and Hatté, C. (2018). Atmosphere–soil carbon transfer as a function of soil depth. Nature, 559(7715):599.

Balesdent, J., Chenu, C., and Balabane, M. (2000). Relationship of soil organic matter dynamics to physical protection and tillage. Soil and tillage research, 53(3-4):215–230.

Barber, D. (2012). Bayesian reasoning and machine learning. Cambridge University Press.

Barbieri, M. M., Berger, J. O., et al. (2004). Optimal predictive model selection. The annals of statistics, 32(3):870–897.

Basile-Doelsch, I., Amundson, R., Stone, W., Masiello, C., Bottero, J. Y., Colin, F., Masin, F., Borschneck, D., and Meunier, J. D. (2005). Mineralogical control of organic carbon dynamics in a volcanic ash soil on la réunion. European Journal of Soil Science, 56(6):689–703.

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

Beal, M. J. et al. (2003). Variational algorithms for approximate Bayesian inference. university of London London.

Becker-Heidmann, P., Andresen, O., Kalmar, D., Scharpenseel, H.-W., and Yaalon, D. H. (2002). Carbon dynamics in vertisols as revealed by high-resolution sampling. Radiocarbon, 44(1):63–73.

Becker-Heidmann, P. and Scharpenseel, H.-W. (1986). Thin layer δ13C and D14C monitoring of "lessive" soil profiles. Radiocarbon, 28(2A):383–390.

Becker-Heidmann, P. and Scharpenseel, H. W. (1989). Carbon isotope dynamics in some tropical soils. Radiocarbon, 31(3):672–679.

Berger, J. O. (2013). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.

Berliner, L. M. (2003). Physical-statistical modeling in geophysics. Journal of Geophysical Research: Atmospheres, 108(D24).

Bol, R., Huang, Y., Meridith, J., Eglinton, G., Harkness, D., and Ineson, P. (1996). The 14C age and residence time of organic matter and its lipid constituents in a stagnohumic gley soil. European Journal of Soil Science, 47(2):215–222.

Bowman, S. and Leese, M. (1995). Radiocarbon calibration-current issues. American Journal of Archaeology, 99(1):102–105.

Box, G. E. (1976). Science and statistics. Journal of the American Statistical Association, 71(356):791–799.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). Handbook of markov chain monte carlo. CRC press.

Brooks, S. P. (2003). Bayesian computation: a statistical revolution. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 361(1813):2681–2697.

Cappé, O., Guillin, A., Marin, J.-M., and Robert, C. P. (2004). Population monte carlo. Journal of Computational and Graphical Statistics, 13(4):907–929.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. Journal of statistical software, 76(1):2–31.

Chen, R.-B., Chu, C.-H., Yuan, S., and Wu, Y. N. (2016). Bayesian sparse group selection. Journal of Computational and Graphical Statistics, 25(3):665–683.

Ciais, P., Sabine, C., Bala, G., Bopp, L., Brovkin, V., Canadell, J., Chhabra, A., DeFries, R., Galloway, J., Heimann, M., et al. (2014). Carbon and other biogeochemical cycles. In Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, pages 465–570. Cambridge University Press.

Clark, J. and Gelfand, A. (2006). Applications of Computational Statistics in the Environmental Sciences: Hierarchical Bayes and MCMC Methods. Oxford University Press.

Coleman, K., Jenkinson, D., Crocker, G., Grace, P., Klir, J., Körschens, M., Poulton, P., and Richter, D. (1997). Simulating trends in soil organic carbon in long-term experiments using rothc-26.3. Geoderma, 81(1-2):29–44.

Conen, F., Leifeld, J., Seth, B., and Alewell, C. (2006). Warming mobilises young and old soil carbon equally. Biogeosciences Discussions, 3(4):1355–1366.

Congdon, P. (2001). Bayesian Statistical Modelling. John Wiley.

Cornuet, J., Marin, J.-M., Mira, A., and Robert, C. P. (2012). Adaptive multiple importance sampling. Scandinavian Journal of Statistics, 39:798–812.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association, 91(434):883–904.

Cressie, N., Calder, C., Clark, J., Ver Hoeff, J., and Wikle, C. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modelling. Ecological Applications, 19(3):553–570.

Currie, L. A. (2004). The remarkable metrological history of radiocarbon dating [ii]. Journal of Research of the National Institute of Standards and Technology, 109(2):185.

Daira, V., François, B., Moreno, B., Xavier, G., and Jean-Michel, L. (2016). Maximum likelihood estimation for a bivariate gaussian process under fixed domain asymptotics. arXiv preprint arXiv:1603.09059.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with nimble. Journal of Computational and Graphical Statistics, 26(2):403–413.

Deckers, J. A. and Nachtergaele, F. (1998). World reference base for soil resources: Introduction, volume 1. Acco.

Dehaene, S. (2011). The number sense: How the mind creates mathematics. OUP USA.

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.

Deng, L., Zhu, G.-y., Tang, Z.-s., and Shangguan, Z.-p. (2016). Global patterns of the effects of land-use changes on soil carbon stocks. Global Ecology and Conservation, 5:127–138.

Donahue, D., Linick, T., and Jull, A. (1990). Isotope-ratio and background corrections for accelerator mass spectrometry radiocarbon measurements. Radiocarbon, 32(2):135–142.

Donnet, S. and Robin, S. (2017). Using deterministic approximations to accelerate smc for posterior sampling. arXiv preprint arXiv:1707.07971.

Doya, K., Ishii, S., Pouget, A., and Rao, R. P. (2007). Bayesian brain: Probabilistic approaches to neural coding. MIT press.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid monte carlo. Physics letters B, 195(2):216–222.

Eclesia, R. P., Jobbagy, E. G., Jackson, R. B., Biganzoli, F., and Piñeiro, G. (2012). Shifts in soil organic carbon for plantation and pasture establishment in native forests and grasslands of South America. Global Change Biology, 18(10):3237–3251.

Falloon, P. and Smith, P. (2010). Modelling soil carbon dynamics, pages 221–244. Cambridge University Press.

Fang, C., Smith, P., Moncrieff, J. B., and Smith, J. U. (2005). Similar response of labile and resistant soil organic matter pools to changes in temperature. Nature, 433(7021):57.

Fierer, N., Craine, J. M., McLauchlan, K., and Schimel, J. P. (2005). Litter quality and the temperature sensitivity of decomposition. Ecology, 86(2):320–326.

Fontaine, S., Mariotti, A., and Abbadie, L. (2003). The priming effect of organic matter: a question of microbial competition? Soil Biology and Biochemistry, 35(6):837–843.

Garcia-Franco, N., Wiesmeier, M., Goberna, M., Martínez-Mena, M., and Albaladejo, J. (2014). Carbon dynamics after afforestation of semiarid shrublands: Implications of site preparation techniques. Forest ecology and management, 319:107–115.

Geiger, R. (1954). Klassifikation der klimate nach w. köppen. Landolt-Börnstein–Zahlenwerte und Funktionen aus Physik, Chemie, Astronomie, Geophysik und Technik, 3:603–607.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. Journal of the American statistical association, 85(410):398–409.

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

Gelman, A. et al. (2013b). Two simple examples for understanding posterior p-values whose distributions are far from uniform. Electronic Journal of Statistics, 7:2595–2602.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. Statistical science, pages 163–185.

Gelman, A., Roberts, G. O., Gilks, W. R., et al. (1996). Efficient metropolis jumping rules. Bayesian statistics, 5:599–608.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. Statistical science, 7(4):457–472.

Geman, S. and Geman, D. (1993). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. Journal of Applied Statistics, 20(5-6):25–62.

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

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. Bayesian statistics, 4:641–649.

Giardina, C. P. and Ryan, M. G. (2000). Evidence that decomposition rates of organic carbon in mineral soil do not vary with temperature. Nature, 404(6780):858.

Hagerty, S. B., Van Groenigen, K. J., Allison, S. D., Hungate, B. A., Schwartz, E., Koch, G. W., Kolka, R. K., and Dijkstra, P. (2014). Accelerated microbial turnover but constant growth efficiency with warming in soil. Nature Climate Change, 4(10):903.

Han, X., Gao, G., Chang, R., Li, Z., Ma, Y., Wang, S., Wang, C., Lü, Y., and Fu, B. (2018). Changes in soil organic and inorganic carbon stocks in deep profiles following cropland abandonment along a precipitation gradient across the loess plateau of china. Agriculture, Ecosystems and Environment, 258:1–13.

Harris, D. C. (2010). Charles david keeling and the story of atmospheric co2 measurements.

Healy, A. F. and Proctor, R. W. (2003). Handbook of psychology: Experimental psychology.

Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J. G., Walsh, M. G., et al. (2014). Soilgrids1km—global soil information based on automated mapping. PLoS One, 9(8):e105992.

Hoff, P. D. (2009). A first course in Bayesian statistical methods. Springer Science & Business Media.

Houghton, J. T. (1995). Climate change 1994: radiative forcing of climate change and an evaluation of the IPCC 1992 IS92 emission scenarios. Cambridge University Press.

Hua, Q., Barbetti, M., and Rakowski, A. Z. (2013). Atmospheric radiocarbon for the period 1950–2010. Radiocarbon, 55(4):2059–2072.

Huang, Y., Li, B., Bryant, C., Bol, R., and Eglinton, G. (1999). Radiocarbon dating of aliphatic hydrocarbons a new approach for dating passive-fraction carbon in soil horizons. Soil Science Society of America Journal, 63(5):1181–1187.

Huang, Y., Roland Bol, D. D. H., and Philip Ineson, G. E. (1996). Post-glacial variations in distributions, 13C and 14C contents of aliphatic hydrocarbons and bulk organic matter in three types of British acid upland soils. Organic Geochemistry, 24(3):273–287.

Hughen, K., Lehman, S., Southon, J., Overpeck, J., Marchal, O., Herring, C., and Turnbull, J. (2004). 14c activity and global carbon cycle changes over the past 50,000 years. Science, 303(5655):202–207.

Jackson, R., Mooney, H., and Schulze, E.-D. (1997). A global budget for fine root biomass, surface area, and nutrient contents. Proceedings of the National Academy of Sciences, 94(14):7362–7366.

Jeffreys, H. (1998). The theory of probability. OUP Oxford.

Jobbágy, E. G. and Jackson, R. B. (2000). The vertical distribution of soil organic carbon and its relation to climate and vegetation. Ecological applications, 10(2):423–436.

Jones, D. L., Olivera-Ardid, S., Klumpp, E., Knief, C., Hill, P. W., Lehndorff, E., and Bol, R. (2018). Moisture activation and carbon use efficiency of soil microbial communities along an aridity gradient in the atacama desert. Soil Biology and Biochemistry, 117:68–71.

Jordan, M., Gahramani, Z., Jaakkola, T., and Saul, K. (1999a). Variational methods for inference and learning in graphical models. Machine Learning, (37):183–233.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999b). An introduction to variational methods for graphical models. Machine learning, 37(2):183–233.

Kadane, J. B. (2011). Principles of uncertainty. Chapman and Hall/CRC.

King, R., Morgan, B., Gimenez, O., and Brooks, S. (2010). Bayesian Analysis for Population Ecology. Chapman & Hall/CRC, Boca Raton, FL.

Köppen, W. (1884). Die wärmezonen der Erde, nach der Dauer der heissen, gemässigten und kalten Zeit und nach der Wirkung der Wärme auf die organische Welt betrachtet. Meteorologische Zeitschrift, 1(21):5–226.

Koppen, W. (1936). Das geographische system der klimat. Handbuch der klimatologie, page 46.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., and Rubel, F. (2006). World map of the köppen-geiger climate classification updated. Meteorologische Zeitschrift, 15(3):259–263.

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

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. (2010). Penalized regression, standard errors, and bayesian lassos. Bayesian Analysis, 5(2):369–411.

Lal, R. (2016). Beyond cop 21: potential and challenges of the "4 per thousand" initiative. Journal of Soil and Water Conservation, 71(1):20A–25A.

Lauritzen, S. and Spiegelhalter, D. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology), 50(2):157–224.

Lee, K. E., Kim, Y., and Xu, R. (2014). Bayesian variable selection under the proportional hazards mixed-effects model. Computational statistics & data analysis, 75:53–65.

Lefevre, R., Barre, P., Moyano, F. E., Christensen, B. T., Bardoux, G., Eglin, T., Girardin, C., Houot, S., Kaetterer, T., Van Oort, F., et al. (2014). Higher temperature sensitivity for stable than for labile soil organic carbon–evidence from incubations of long-term bare fallow soils. Global Change Biology, 20(2):633–640.

Legros, J.-P. (2007). Les grands sols du monde. PPUR presses polytechniques.

Lele, S. R., Dennis, B., and Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using bayesian markov chain monte carlo methods. Ecology letters, 10(7):551–563.

Li, Y., Zhang, J., Chang, S. X., Jiang, P., Zhou, G., Shen, Z., Wu, J., Lin, L., Wang, Z., and Shen, M. (2014). Converting native shrub forests to chinese chestnut plantations and subsequent intensive management affected soil C and N pools. Forest ecology and management, 312:161–169.

Libby, W. F., Anderson, E. C., and Arnold, J. R. (1949). Age determination by radiocarbon content: world-wide assay of natural radiocarbon. Science, 109(2827):227–228.

Lindley, D. V. (2013). Understanding uncertainty. John Wiley & Sons.

Liu, J. S., Chen, R., and Logvinenko, T. (2001). A theoretical framework for sequential importance sampling with resampling. In Sequential Monte Carlo methods in practice, pages 225–246. Springer.

Liu, Z., Shao, M., and Wang, Y. (2011). Effect of environmental factors on regional soil organic carbon stocks across the loess plateau region, china. Agriculture, Ecosystems & Environment, 142(3-4):184–194.

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs - a Bayesian modelling framework: Concepts, structure, and extensibility. Statistics and Computing, 10(4):325–337.

Lützow, M. v., Kögel-Knabner, I., Ekschmitt, K., Matzner, E., Guggenberger, G., Marschner, B., and Flessa, H. (2006). Stabilization of organic matter in temperate soils: mechanisms and their relevance under different soil conditions–a review. European Journal of Soil Science, 57(4):426–445.

Maestre, F. T., Delgado-Baquerizo, M., Jeffries, T. C., Eldridge, D. J., Ochoa, V., Gozalo, B., Quero, J. L., Garcia-Gomez, M., Gallardo, A., Ulrich, W., et al. (2015). Increasing aridity reduces soil microbial diversity and abundance in global drylands. Proceedings of the National Academy of Sciences, 112(51):15684–15689.

Malsiner-Walli, G., Pauger, D., and Wagner, H. (2017). Effect fusion using model-based clustering. Statistical Modelling, page 1471082X17739058.

Malsiner-Walli, G., Pauger, D., and Wagner, H. (2018). Effect fusion using model-based clustering. Statistical Modelling, 18(2):175–196.

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.

Mazaud, A., Laj, C., Bard, E., Arnold, M., and Tric, E. (1992). A geomagnetic calibration of the radiocarbon time-scale. Bard E, Broecker WS, pages 163–175.

McBratney, A. B., Santos, M. M., and Minasny, B. (2003). On digital soil mapping. Geoderma, 117(1-2):3–52.

McGrayne, S. B. (2011). The Theory That Would Not Die. Yale University Press, New Haven, CT.

Monnahan, C. C., Thorson, J. T., and Branch, T. A. (2017). Faster estimation of bayesian models in ecology using hamiltonian monte carlo. Methods in Ecology and Evolution, 8(3):339–348.

Naganawa, T., Kyuma, K., Yamamoto, H., Yamamoto, Y., Yokoi, H., and Tatsuyama, K. (1989). Measurement of soil respiration in the field: influence of temperature, moisture level, and application of sewage sludge compost and agro-chemicals. Soil science and plant nutrition, 35(4):509–516.

Nave, L., Swanston, C., Mishra, U., and Nadelhoffer, K. (2013). Afforestation effects on soil carbon storage in the united states: a synthesis. Soil Science Society of America Journal, 77(3):1035–1047.

Neal, R. M. (2011). Mcmc using hamiltonian dynamics. Chapter 5 of Handbook of Markov Chain Monte Carlo, edited by S. Brooks, A. Gelman, G. Jones and X. Meng, pages 113–162.

New, M., Lister, D., Hulme, M., and Makin, I. (2002). A high-resolution data set of surface climate over global land areas. Climate research, 21(1):1–25.

Noojipady, P., Morton, C. D., Macedo, N. M., Victoria, C. D., Huang, C., Gibbs, K. H., and Bolfe, L. E. (2017). Forest carbon emissions from cropland expansion in the brazilian cerrado biome. Environmental Research Letters, 12(2):025004.

Ntzoufras, I. (2011). Bayesian modeling using WinBUGS, volume 698. John Wiley & Sons.

Oades, J. (1988). The retention of organic matter in soils. Biogeochemistry, 5(1):35–70.

Parent, E. and Rivot, E. (2012). Introduction to hierarchical Bayesian modeling for ecological data. Chapman and Hall/CRC.

Parisi, G. (1988). Statistical field theory. Addison-Wesley.

Parton, W., Schimel, D. S., Cole, C., and Ojima, D. (1987). Analysis of factors controlling soil organic matter levels in great plains grasslands. Soil Science Society of America Journal, 51(5):1173–1179.

Paul, E. A. (2016). The nature and dynamics of soil organic matter: Plant inputs, microbial transformations, and organic matter stabilization. Soil Biology and Biochemistry, 98:109–126.

Paul, S., Flessa, H., Veldkamp, E., and López-Ulloa, M. (2008). Stabilization of recent soil carbon in the humid tropics following land use changes: evidence from aggregate fractionation and stable isotope analyses. Biogeochemistry, 87(3):247–263.

Plummer, M. et al. (2003). Jags: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd international workshop on distributed statistical computing, number 125.10. Vienna, Austria.

Prather, M. J., Holmes, C. D., and Hsu, J. (2012). Reactive greenhouse gas scenarios: Systematic exploration of uncertainties and the role of atmospheric chemistry. Geophysical Research Letters, 39(9).

Prentice, I. C., Farquhar, G., Fasham, M., Goulden, M., Heimann, M., Jaramillo, V., Kheshgi, H., LeQuéré, C., Scholes, R., and Wallace, D. W. (2001). The carbon cycle and atmospheric carbon dioxide. Cambridge University Press.

Rabbi, S., Tighe, M., Delgado-Baquerizo, M., Cowie, A., Robertson, F., Dalal, R., Page, K., Crawford, D., Wilson, B. R., Schwenke, G., et al. (2015). Climate and soil properties limit the positive effects of land use reversion on carbon storage in eastern australia. Scientific Reports, 5:17866.

Raftery, A. E. and Lewis, S. M. (1992). [practical markov chain monte carlo]: comment: one long run with diagnostics: implementation strategies for markov chain monte carlo. Statistical science, 7(4):493–497.

Raftery, A. E., Zimmer, A., Frierson, D. M. W., Startz, R., and Liu, P. (2017). Less than 2 °c warming by 2100 unlikely. Nature Climate Change, 7:634–641.

Rasmussen, C. E. (2004). Gaussian processes in machine learning. In Advanced lectures on machine learning, pages 63–71. Springer.

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

Reimer, P. J., Bard, E., Bayliss, A., Beck, J. W., Blackwell, P. G., Ramsey, C. B., Buck, C. E., Cheng, H., Edwards, R. L., Friedrich, M., et al. (2013). Intcal13 and marine13 radiocarbon age calibration curves 0–50,000 years cal bp. Radiocarbon, 55(4):1869–1887.

Reimer, P. J., Brown, T. A., and Reimer, R. W. (2004). Discussion: reporting and calibration of post-bomb 14c data. Radiocarbon, 46(3):1299–1304.

Ren, C., Chen, J., Lu, X., Doughty, R., Zhao, F., Zhong, Z., Han, X., Yang, G., Feng, Y., and Ren, G. (2018). Responses of soil total microbial biomass and community compositions to rainfall reductions. Soil Biology and Biochemistry, 116:4–10.

Robert, C. (1992). L'analyse statistique bayésienne. Economica.

Robert, C. P. (2004). Monte carlo methods. Wiley Online Library.

Roberts, G. (1996). Markov Chain Concepts related to sampling algorithms, pages 45–58. Chapman and Hall, London.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. The Annals of Applied Probability, 7:110–120.

Rosenthal, J. S. (2014). Optimising and adapting the metropolis algorithm. Chapter 6 of Statistics in Action : a Canadian Outlook edited by Jerald F. Lawless, CRC Press, pages 93–108.

Smith, P., Fang, C., Dawson, J. J., and Moncrieff, J. B. (2008). Impact of global warming on soil organic carbon. Advances in agronomy, 97:1–43.

Song, X.-P., Hansen, M. C., Stehman, S. V., Potapov, P. V., Tyukavina, A., Vermote, E. F., and Townshend, J. R. (2018). Global land change from 1982 to 2016. Nature, 560(7720):639.

Spaargaren, O. (2001). Major soils of the world. Wageninen, The Netherlands: International Soil Reference and Information Centre.

Spiegelhalter, D., Thomas, A., and Best, N. (2003). Winbugs version 14 user manual. MRC and Imperial College of Science, Technology and Medicine, (available at: http://www.mrc-bsu.cam.ac.uk/bugs).

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.

Stocker, T., Qin, D., Plattner, G., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. (2013). Ipcc, 2013: Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change, 1535 pp.

Stone, L. D., Keller, C. M., Kratzke, T. M., and Strumpfer, J. P. (2011). Search analysis for the underwater wreckage of air france flight 447. In Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on, pages 1–8. IEEE.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the royal statistical society. Series B (Methodological), pages 111–147.

Stuiver, M. and Quay, P. (1981). Atmospheric14c changes resulting from fossil fuel co2 release and cosmic ray flux variability. Earth and Planetary Science Letters, 53(3):349–362.

Stuiver, M. and Reimer, P. J. (1986). A computer program for radiocarbon age calibration. Radiocarbon, 28(2B):1022–1030.

Suess, H. E. (1955). Radiocarbon concentration in modern wood. Science, 122(3166):415–417.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288.

Tifafi, M., Guenet, B., and Hatté, C. (2018). Large differences in global and regional total soil carbon stock estimates based on SoilGrids, HWSD, and NCSCD: Intercomparison and evaluation based on field data from Usa, England, Wales, and France. Global Biogeochemical Cycles, 32(1):42–56.

Todd-Brown, K., Randerson, J., Post, W., Hoffman, F., Tarnocai, C., Schuur, E., and Allison, S. (2013). Causes of variation in soil carbon simulations from cmip5 earth system models and comparison with observations. Biogeosciences, 10(3):1717–1736.

Trumbore, S. E., Chadwick, O. A., and Amundson, R. (1996). Rapid exchange between soil carbon and atmospheric carbon dioxide driven by temperature change. Science, 272(5260):393–396.

Vehtari, A., Gelman, A., and Gabry, J. (2015). Efficient implementation of leave-one-out cross-validation and waic for evaluating fitted bayesian models. arXiv preprint arXiv:1507.04544.

Wang, Y., Li, Y., Ye, X., Chu, Y., and Wang, X. (2010). Profile storage of organic/inorganic carbon in soil: From forest to desert. Science of the Total Environment, 408(8):1925–1931.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research, 11(Dec):3571–3594.

Wiesmeier, M., Dick, D., Rumpel, C., Dalmolin, R., Hilscher, A., and Knicker, H. (2009). Depletion of soil organic carbon and nitrogen under Pinus taeda plantations in southern Brazilian grasslands (Campos). European journal of soil science, 60(3):347–359.

Wikle, C. (2003). Hierarchical models in environmental science. International Statistical Review, 71(2):181–199.

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

Yan, D., Li, J., Pei, J., Cui, J., Nie, M., and Fang, C. (2017). The temperature sensitivity of soil organic carbon decomposition is greater in subsoil than in topsoil during laboratory incubation. Scientific Reports, 7(1):5181.

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

Zhou, G., Zhou, C., Liu, S., Tang, X., Ouyang, X., Zhang, D., Liu, S., Liu, J., Yan, J., Zhou, C., et al. (2006). Belowground carbon balance and carbon accumulation rate in the successional series of monsoon evergreen broad-leaved forest. Science in China Series D, 49(3):311–321.

Zomer, R., Trabucco, A., van Straaten, O., and Bossio, D. (2006). Carbon, land and water: A global analysis of the hydrologic dimensions of climate change mitigation through afforestation/reforestation, volume 101. IWMI.

Zomer, R. J., Trabucco, A., Bossio, D. A., and Verchot, L. V. (2008). Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. Agriculture, ecosystems & environment, 126(1-2):67–80.

Zuur, A. F., Iono, E. N., and Walker, J. N. (2009). Mixed Effect Models and Extension in Ecology with R. Springer, New York, NY.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences de l'environnement d'île-de-France (SEIF)

Titre : Dynamique verticale du carbone dans les sols – Utilisation combinée des traceurs isotopiques et de méta-analyse statistique.

Mots clés : carbone ; géochimie, modélisation probabiliste ; statistique bayésienne ; méta-analyses

Résumé : Bien qu'il s'agisse du plus grand réservoir terrestre interagissant avec l'atmosphère, la réponse du carbone du sol au changement climatique et à l'évolution de l'utilisation des terres demeure incertaine. Pour mieux comprendre la dynamique du carbone du sol et évaluer l'impact des facteurs climatiques et environnementaux sur les stocks et le temps moyen de résidence du carbone du sol, un modèle non linéaire hiérarchique d'effets aléatoires a été proposé pour modéliser la variation des réponses du carbone en fonction de la profondeur. La sélection des facteurs climatiques et environnementaux a reposé sur trois techniques de sélection bayésienne (Bayesian Group Lasso, Bayesian Sparse Group Selection et Bayesian Effcet Fusion) appropriées pour les prédicteurs catégoriels (type de sol et type d'écosystème) et sur le Stochastic Search Variable Selection pour les prédicteurs numériques (température, précipitations, etc.). La modélisation statistique a également permis d'étudier l'effet de l'augmentation de la température et de la conversion de l'utilisation des terres sur la dynamique du carbone du sol.

Title : Vertical dynamics of soil carbon - Combined use of isotopic tracers and statistical meta-analysis

Keywords : carbon ; geochemistry ; probabilistic modeling ; bayesian approach ; meta-analysis

Abstract : Although it is the largest land reservoir interacting with the atmosphere, the response of the soil carbon reservoir to climate change and land use evolution remains uncertain. To better understand the soil carbon dynamics and assess the impact of. climate and environmental factors on residence time and soil organic carbon content, a non-linear hierarchical random effects model was proposed to model the variation in the responses of soil carbon content and soil radiocarbon as a function of depth. The selection of climatic and environmental factors was based on three Bayesian Selection techniques (Bayesian Group Lasso, Bayesian Sparse Group Selection and Bayesian Effect fusion) appropriate for categorical predictors (soil type and ecosystem type) and on the Stochastic Search Variable Selection for the numerical predictors (temperature, precipitation, etc.). The statistical modeling also enabled the effect of temperature increase and land-use conversion on soil carbon dynamics to be investigated

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France