

Compte rendu de la 1^{er} année de thèse

Rana JREICH

sous la direction de Christine HATTE et Eric PARENT

Analyses statistiques de la distribution verticale du $\Delta^{14}\text{C}$ dans les sols



1^{er} novembre 2015 — 3 Août 2016

Table des matières

1	Introduction Générale	4
1.1	Intérêt de l'estimation du stock mondial de carbone	4
1.2	Court résumé sur l'accumulation du carbone dans le sol et lien avec le $\Delta^{14}C$	6
2	Description de la base de données	15
3	Modélisation statistique des profils $\Delta^{14}C$	21
4	Approches statistiques	24
4.1	Régression non-linéaire pour estimer les variables latentes-caractéristique de forme- $\phi_1, \phi_2, \phi_3, \phi_4$	24
5	Variabilité intra-site	25
5.1	Variabilité intersite	26
6	Lecture et Nettoyage des données	27
7	Inférence bayésienne sous Jags	27
8	Tourner le modèle sur des données simulées	30
8.1	Diagnostic de la convergance sur les données simulées	30
8.2	Comparaison entre la matrice Θ_{true} et $\Theta_{posterior}$	33
9	Application finale sur les données réelles de $\Delta^{14}C$	33
10	Sélection du modèle dans le cadre bayésien	33
11	Perspectives	41
A	Les profils sélectionnés pour l'analyse statistique	41

B	Code R : Estimation via "optim"	48
C	Code R : Nettoyage des données	60
D	Code R : Manova et données simulées	66
E	Code R : Modèle sous Jags	74
F	Diagnostic de la convergence	74
G	Test du modèle sous Jags	75
H	Code R : Application sur les données réelles	76
I	Code R : Sélection de variable	85

Table des figures

1	Cycle de carbone terrestre.	7
2	Répartition de la matière organique dans le sol.	8
3	Les Processus de protection de la matière organique dans le sol.	9
4	Différence du rapport isotopique $\delta^{13}C$ entre les plantes C3 et C4.	11
5	Profils de carbone avant et après la monoculture de plante de type photosynthétique différent.	12
6	Augmentation de la concentration du carbone dans l'atmosphère suite aux essais nucléaires.	13
7	Profils de carbone avant et après le pic de bombes.	14
8	Répartition géographique des sites échantillonés.	15
9	Description de la base de données.	16

10	Profondeur maximale des profils en fonction de types de sols (en abscisse), La courbe rouge en pointillé délimite les profils qui vont être sélectionnés pour l'estimation finale (ce sont ceux dont la mesure la plus profonde se situe suite à droite de la courbe rouge. Il en résulte une sélection de 159 profils.	19
11	Répartition du nbr de sites par type de sol.	20
12	Répartition du nbr de sites par type d'écosystème.	21
13	Profils de carbone pour tous les sites étudiés.	22
14	Valeurs et histogarammes des variables- ou caractéristiques de forme - estimées via "optim" ϕ_1, ϕ_2, ϕ_3 et ϕ_4 pour l'enssemble des 159 sites de l'étude.	25
15	Les quantités aléatoires sont entourées par des ellipses et les quantités fixes ou observées sont entourées par des rectangles.	28
16	Probabilité de sélection <i>a posteriori</i> dans le modèle pour chacune des variables latentes - caractéristiques du modèle.	40

Liste des tableaux

1	Nombre de profils par type de sol, pour les 159 utilisés. Certaines catégories contiennent d'autres types de sols que celui indiqué en titre : "luvisol" contient les types "planosol" et "phaaeozem", et "ferrasol" contient le type "plinthosol".	20
2	Nombre de profils par type d'écosystème, pour les 159 utilisés. NA représente les profils pour lesquels ni le type de végétation ni l'usage des terres ne sont renseignés.	21
3	Représentation graphique du modèle choisi.	23

1 Introduction Générale

1.1 Intérêt de l'estimation du stock mondial de carbone

Un des défis majeur qui domine notre époque, auquel doivent répondre les organismes de contrôle environnementaux, est le **réchauffement climatique**.

Ce réchauffement est dû à l'émission des gaz à effet de serre d'origine anthropique principalement le CO_2 et le méthane.

Divers changements observés conduisent à la conclusion de l'existence d'un réchauffement climatique comme : La réduction de la superficie des glacières, effet géophysiques et sismiques, extinctions d'espèces, le recul de la banquise et d'autres...

Un autre impact très inquiétant concerne le dégel de Permafrost ou Pergélisol. C'est quoi le Permafrost ? Le Permafrost est un sol dont la température se maintient au dessous de $0^{\circ}C$ pendant plus de 2 ans consécutifs. Il représente 20 % de la surface de planète.

En effet cette couche gelée, depuis des millions d'années, renferme 1700 milliards de tonnes de carbone soit le double du CO_2 atmosphérique.

Ainsi le processus de dégel pourrait contribuer à libérer des milliards de tonnes dans l'atmosphère, ainsi plus de gaz à effet de serre, signifie que le réchauffement pourrait être pire que prévu.

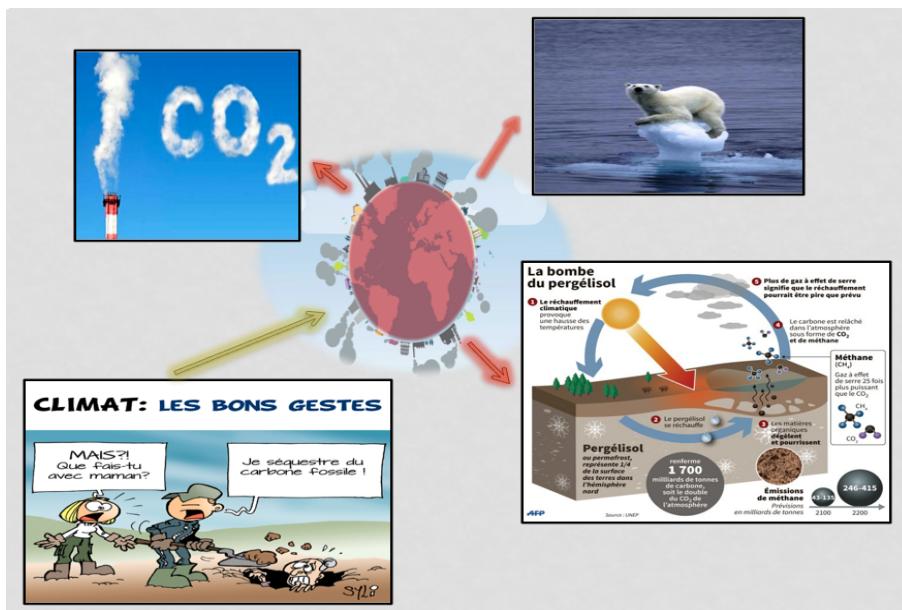
On peut voir donc que la situation est très inquiétante et les conséquences immédiates pèsent aussi sur les générations à venir.

Raison pour laquelle il est essentiel de mettre en place plusieurs actions afin de permettre d'épargner des vies à le long terme.

Face au défi climatique, l'accord reconnaît une responsabilité partagée mais différenciée des Etats, c'est-à-dire en fonction des capacités respectives et des contextes nationaux différents. En 2015 la 21e conférence des parties à la Convention-cadre des Nations unies sur les changements climatiques a eu lieu à Paris. Chaque année, les participants de cette conférence se réunissent pour décider des mesures à mettre en place, cette fois la limitation du réchauffement climatique mondiale entre 1.5 et 2 $0^{\circ}C$ d'ici 2100. L'une des solutions proposées dans le protocole de Kyoto afin réduire les émission des GES était d'augmenter la séquestration du carbone dans le sol.

C'est pourquoi ma thèse a pour intérêt d'évaluer la quantité de carbone dans le sol au niveau mondial, en fonction des divers scénarios comme le changement climatique et le changement d'usage du sol.

Ceci met en évidence l'importance de comprendre la dynamique du carbone dans le sol surtout que le GIEC (groupe d'experts intergouvernementale sur l'évolution du climat) a mis en évidence de grandes incertitudes sur les estimations futures du cycle global du carbone, en particulier, la réponse du carbone organique des sols face aux changements de climat et d'usage des terres.



Une incertitude forte est portée sur le carbone profond surtout que la plupart des modèles sur la dynamique de carbone sont essentiellement calés à partir des études de carbone pour les premiers 20 ou 30 centimètres, ils prennent mal en compte le carbone profond.

Mon travail de thèse a pour objectif de répondre à plusieurs questions :

1. D'abord, estimer la quantité globale du stock de carbone dans le sol au niveau mondial.
2. Identifier les facteurs environnementaux (température, Précipitation, type de sol etc..) qui influent le plus sur la dynamique de carbone.
3. Essayer d'avoir une vision sur la modification de ce stock en cas du réchauffement climatique et changement d'usage de sols.
4. Améliorer le plan expérimentale des données afin d'améliorer la précision.

1.2 Court résumé sur l'accumulation du carbone dans le sol et lien avec le $\Delta^{14}C$

Cycle de carbone : il existe 4 réservoirs de carbone : l'hydrosphère, la lithosphère, la biosphère et l'atmosphère.

La plus grande partie du carbone terrestre est piégé dans des composés qui participent peu au cycle de carbone comme les roches sédimentaires et les masses d'eaux océanique profondes qui piègent respectivement 50 000 et 39 040 gigatonnes (Gt).

Pourquoi s'intéresse-t-on plutôt au carbone du sol ? le carbone organique du sol représente le plus grand réservoir en interaction avec l'atmosphère, il est estimé entre 1700 et 2000 Gt à 1m de profondeur ce qui est équivalent à 231 années de combustion fossile. En plus, le sol peut être considéré comme puit ou source du carbone, ainsi si le flux est positif on parle de séquestration et s'il est négatif, on parle d'émission.

La végétation piège 650 Gt alors que l'atmosphère 750 Gt, emmagasinant considérablement moins que le sol.

La végétation terrestre piège du carbone d'origine atmosphérique à travers le processus de la production primaire (la photosynthèse). Le carbone par la suite sera envoyé vers le sol, soit à travers ses racines, soit sous forme de matières organiques mortes.

Une grande partie de ce carbone est restitué dans l'atmosphère par des processus de respiration et de décomposition de la matière organique dans le sol. Ce flux est estimé par 120 Gt.

En outre, le flux naturel entre l'atmosphère et l'hydrosphère est exprimé d'un part par la forte solubilité du CO_2 dans l'eau des océans et d'autre part par la respiration des êtres vivants marins.

Ce flux estimé par 180 Gt dépend de plusieurs facteurs tel que la température de l'eau des océans (les eaux froides contiennent plus de CO_2 dissous que les eaux chaudes).

Dans le passé, le développement de l'agriculture a été la cause principale de l'augmentation du CO_2 dans l'atmosphère, mais actuellement la combustion du carbone fossile par l'industrie et les transports représente la contribution principale et libère un flux de 6.5 Gt dans l'atmosphère. Ainsi le déboisement enlève un puits potentiel du CO_2 et libère 1.6 Gt. Environ la moitié de ce carbone est réabsorbée par la biosphère et les océans et ceci est du aux processus de respiration et de dissolution.

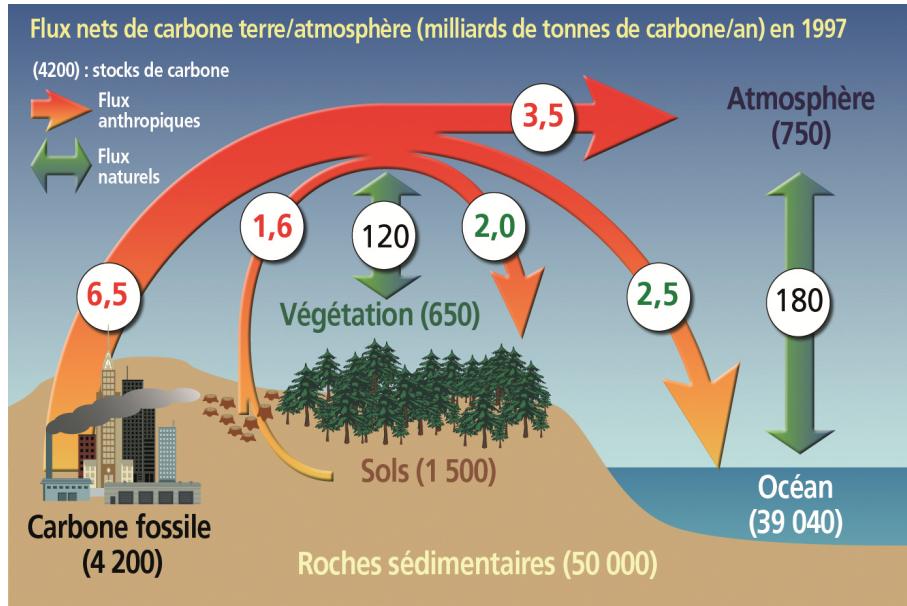


FIGURE 1 – Cycle de carbone terrestre.

Le cycle de carbone sera-t-il affecté au cas du réchauffement climatique surtout au niveau du sol ? Que se passe t'il si on remplace la forêt par une prairie ? Est ce que le sol se comportera comme un puit ou comme une source d'émission de carbone ? Quelques questions auxquelles on cherche à trouver des réponses.

Répartition de la matière organique dans le sol : Le sol est représenté sous forme de strates plus ou moins bien définies et continues, le sol est donc caractérisé par cette structure complexe. En pratique la matière organique est principalement présentée dans la partie superficielle du sol et sa concentration diminue en fonction de le profondeur.

Le stock de carbone du sol est défini d'un part par un apport de flux entrants au sol et d'autre part par des vitesses de minéralisation.

Dans le sol, la minéralisation est assurée par la décomposition microbienne de la matière organique. Cette décomposition libère des éléments nutritifs nécessaires à la croissance des plantes et du CO_2 dans l'atmosphère. Ainsi on peut imaginer la répartition du carbone en 3 compartiments :

Le premier très labile, biodégradable entre 1-5 ans, il recueille 75% des apports

annuels du sol. Le reste peut être répartie en 2 compartiments : l'un caractérisé par un taux de décomposition très lent et dont le temps moyenne de résidence est de 25 ans, et l'autre Stable dont le temps de résidence est de plus de 1000 ans. Le défaut principale est que la méconnaissance des processus gérant la dynamique de carbone dans les sols proviennent de l'absence de répartition des compartiments, par exemple, un carbone qui est piégé longtemps dans le sol et qui se trouve dans l'état passif peut recontribuer au cycle de carbone à cause du changement de type de végétation ou d'usage de sol.

Les flux rentrants sont largement maîtrisés par l'homme. En revanche, les vitesses de minéralisation sont variables et moins maîtrisées, raison pour laquelle la recherche s'est principalement focalisée sur la dynamique du carbone dans les sols.

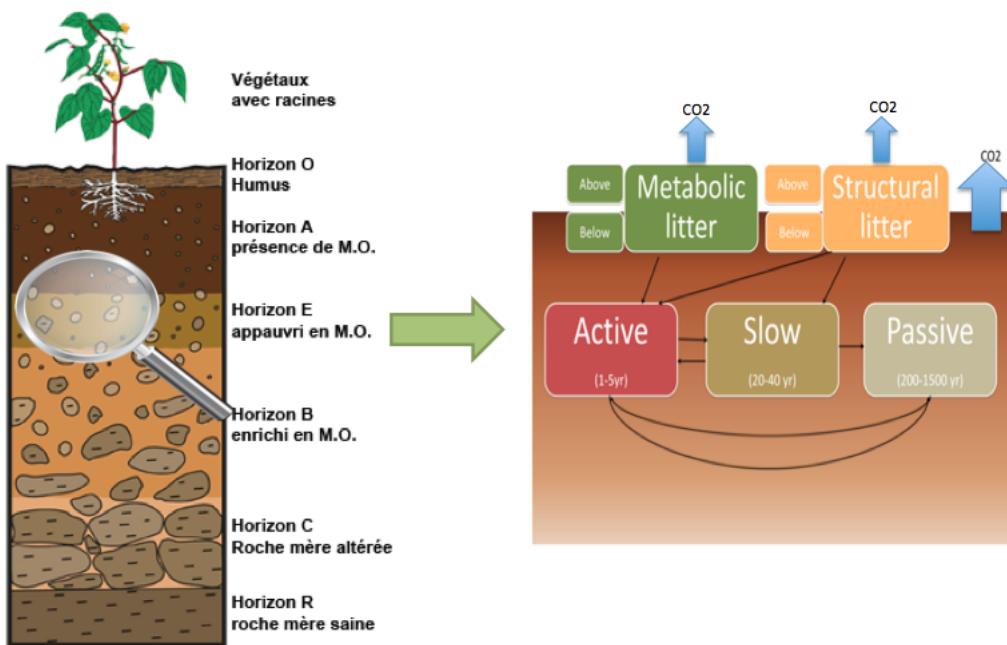


FIGURE 2 – Répartition de la matière organique dans le sol.

L'une des demandes du Protocole de Kyoto en 1992 était la compréhension fondamentale du carbone dans le sol.

Or, il existe une grande incertitude sur les mécanismes et les processus qui peuvent ralentir le processus de minéralisation et protéger le carbone assez longtemps (Inaccessibilité spatiale, Hydrophobie, Occlusion de la matière organique par agrégation).



FIGURE 3 – Les Processus de protection de la matière organique dans le sol.

On a également vu que certains processus avaient plus de poids que d'autres. Mais dans l'ensemble, on ne sait pas hiérarchiser ces processus et on ne connaît pas leur poids relatifs.

Cela nous conduit à étudier le carbone à l'échelle globale, en intégrant tous ces processus, et en reliant la résultante aux conditions environnementales.

Le modèle statistique à construire se place en parallèle de la compréhension des processus, et permet d'apporter des réponses aux questions immédiates pour aujourd'hui et pour un futur proche.

Description des techniques de traçage isotopique du carbone dans le sol :

A l'origine, les données de 200 sites réparties sur tout le globe terrestre ont été collectées à partir d'articles de la littérature. Mon travail consiste à interpoler les quantités du carbone à partir de mesures de ^{13}C et ^{14}C à différents niveau de profondeur.

Quelles sont les techniques utilisées afin de tracer la dynamique du carbone dans le sol ?

Afin de représenter la matière organique (M.O), de la spécifier, de suivre et de donner une cinétique, des méthodes de traçage isotopiques ont été développées telles que le traçage isotopique en ^{13}C . Cette technique exige un changement de végétation en fonction des différents types photosynthétiques, tandis que la datation en ^{14}C est liée, à l'introduction massive de ^{14}C suite aux explosions nucléaires ou au ^{14}C naturel.

La différence entre ces 2 techniques est que la première peut être utilisée pour des temps de résidence allant d'une année à un siècle alors que la deuxième peut être utilisée pour des temps de résidence allant d'une année à plusieurs millénaires.

Abondance naturelle en ^{13}C : Le carbone a 2 isotopes stables : le ^{12}C et le ^{13}C . Le premier se trouve en abondance naturelle de 98.93% alors que le second est de 1.1%.

Ainsi, on définit le rapport isotopique comme le rapport entre le ^{13}C et le ^{12}C . Pour cette technique, on définit le critère δ comme étant la différence relative des rapports isotopiques de l'échantillon à la référence qui est le PDB.

$$\delta^{13}\text{C}(\text{\% vs PDB}) = \left[\frac{\frac{^{13}\text{C}}{^{12}\text{C}}_{\text{Échantillon}} - \frac{^{13}\text{C}}{^{12}\text{C}}_{\text{PDB}}}{\frac{^{13}\text{C}}{^{12}\text{C}}_{\text{PDB}}} \right] \times 1000$$

Si $\delta > 0$, il y a plus de carbone lourds dans l'échantillon que dans la référence. La quasi-totalité des plantes continentales des pays tempérées et froides emploient le cycle photosynthétique de type 3.

La photosynthèse en C4 est découverte en 1996, cette adaptation est apparue chez nombreux groupes de plantes, principalement comme adaptation au stress hydrique ou à une réduction de disponibilité de CO_2 pendant la journée. Ces derniers sont essentiellement des plantes tropicales et certains graminés comme maïs et canne à sucre.

Au niveau du fractionnement isotopique, les plantes en C4 sont plus riches en

carbone ^{13}C que les plantes en C3.

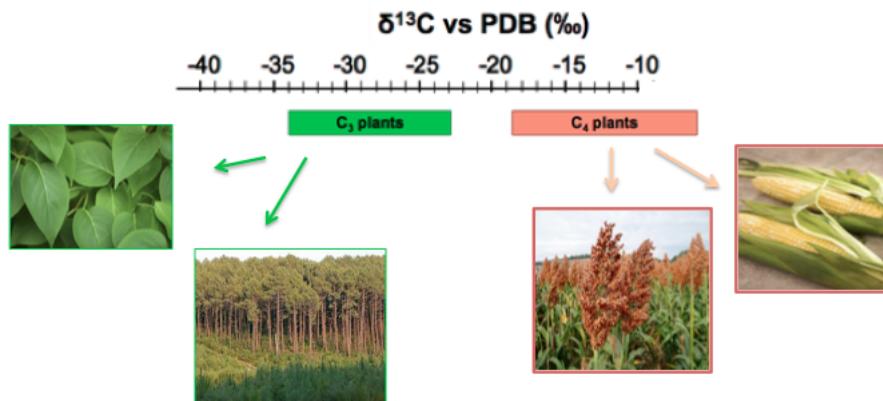


FIGURE 4 – Différence du rapport isotopique $\delta^{13}\text{C}$ entre les plantes C3 et C4.

Quel est l'intérêt de changer d'un type de végétation à un autre ?

La composition isotopique de la matière organique du sol est très proche de celle de la végétation en équilibre, ainsi la monoculture des plantes C4 sur les sols entièrement occupés par une végétation en C3 ou l'inverse est un excellent marquage de la matière organique incorporée dans le sol.

En effet si le sol était entièrement occupé par des arbres de type C3, son profil de carbone correspond à une dynamique verticale et la valeur de $\delta^{13}\text{C}$ est autour de 26 ‰. Suite à une monoculture de maïs, et vu que les C4 sont plus riches en ^{13}C , celle-ci conduit à un enrichissement isotopique en ^{13}C car la couche superficielle du sol possède à peu près le même fractionnement isotopique des plantes C4 à la surface, qui est autour de -12 ‰. Voir la figure ci-dessous :

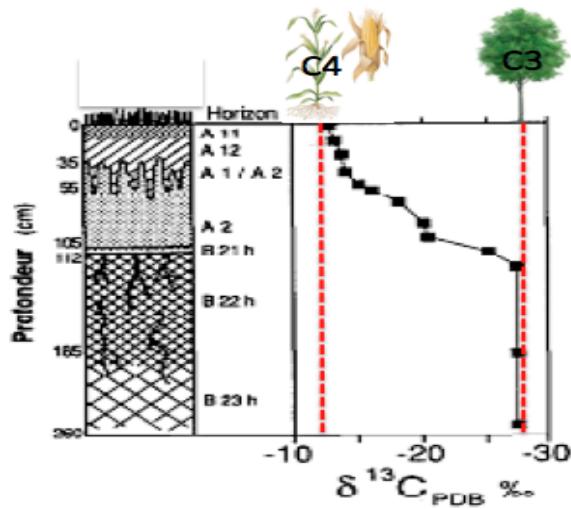


FIGURE 5 – Profils de carbone avant et après la monoculture de plante de type photosynthétique différent.

Datation par le ^{14}C : Le carbone 14 est radioactif et son abondance naturelle est de l’ordre de 10^{-12} , sa période de décroissance est de 5730 ans c-à-d qu’il perd la moitié de sa quantité tous les 5730 ans.

La méthode suppose que la végétation est intégrée d’une proportion constante de ^{14}C naturelle avec le temps .

En sciences du sol, il est bien difficile d’apprécier l’incidence des variations naturelles de production du ^{14}C mais il existe cette fois-ci une variation artificielle, à laquelle le sol ne s’échappe pas, il s’agit de l’introduction récente et massive de ^{14}C dans l’atmosphère suite aux explosions nucléaires depuis 1950.

La figure ci-dessous nous permet de voir les écarts de concentration de ^{14}C par rapport au niveau naturel (représenté par la droite horizontale).

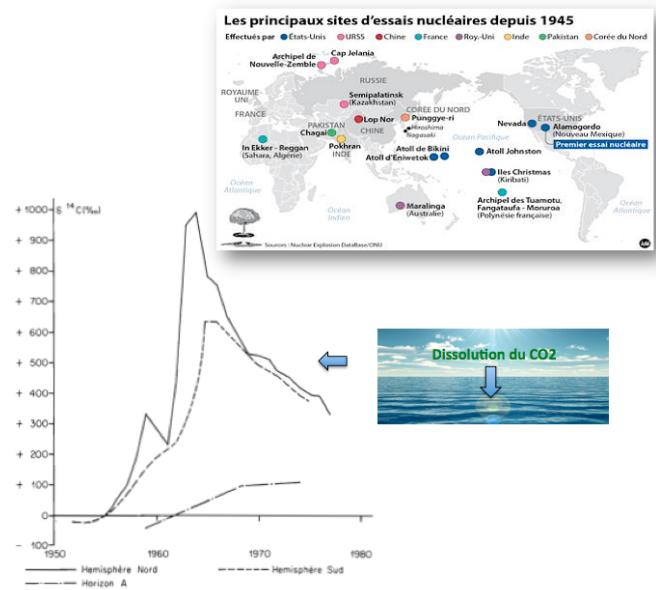


FIGURE 6 – Augmentation de la concentration du carbone dans l’atmosphère suite aux essais nucléaires.

Les teneurs en ^{14}C de l’atmosphère ont doublé en 1962 par rapport au niveau naturel suivi d’une diminution à la suite, due à la dilution du CO_2 dans le réservoir océanique.

Ainsi, comme les végétaux ont subi les mêmes variations de l’atmosphère, leurs résidus humifiés se trouvent contaminées par le ^{14}C . La figure ci-dessous, nous permet de mieux comprendre la différence au niveau de la dynamique du carbone du sol avant et après le pic des bombes :

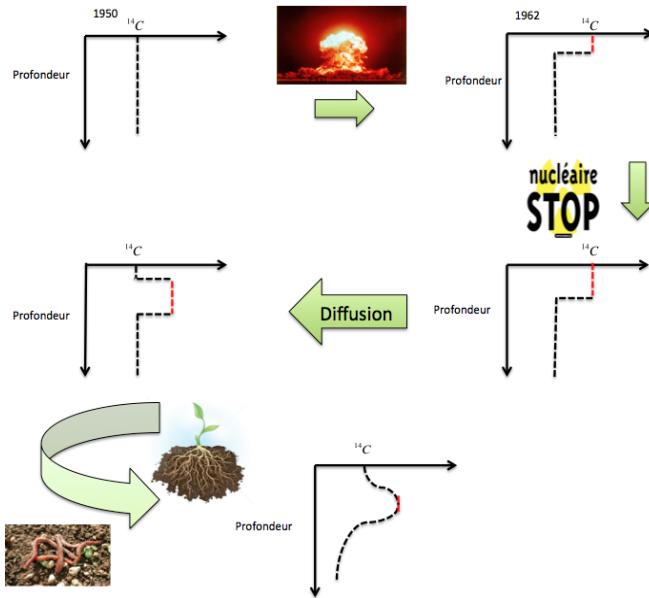


FIGURE 7 – Profils de carbone avant et après le pic de bombes.

Avant le pic des bombes, la proportion du carbone 14 est intégrée dans le sol avec une proportion constante, ainsi, après les essais nucléaires et l'introduction massive du carbone 14, on remarque que la proportion de carbone 14 intégrée dans la surface a augmenté.

Une fois les essais nucléaires sont interdits, le processus de diffusion de la matière organique est responsable de la descente de la bosse reliée au pic, et finalement, grâce aux racines et aux processus de la bioturbation, cette bosse devient de plus en plus arrondie.

Ainsi le dernier profil peut être considéré comme le profil type de la dynamique du ^{14}C dans le sol.

Pour conclure, le principe de la datation par le ^{14}C est basé sur la loi de décroissance radioactive.

$$A = A_0 * \exp(-\lambda t) \quad \lambda = \frac{\ln 2}{T}$$

avec

- A_0 : l'activité spécifique initiale. On pourra la définir comme étant celle du carbone des organismes vivants.

- A : l'activité spécifique de l'échantillon à dater.
- T : la période de radioactivité fixée à 5730 ans.

Par conséquent l'âge t de l'échantillon se déduit de la formule suivante.

$$t = 8.035 \ln \frac{A_0}{A}$$

2 Description de la base de données

A l'origine, les données de 200 sites ont été obtenues à partir de la littérature : elles décrivent des profils échantillonés, certains datent de plus de 80 ans. Cependant toutes ces données ne sont pas utilisables dans notre étude.

Ces sites sont des fosses distribuées à peu près partout sur le globe terrestre (voir la figure ci-dessous) :

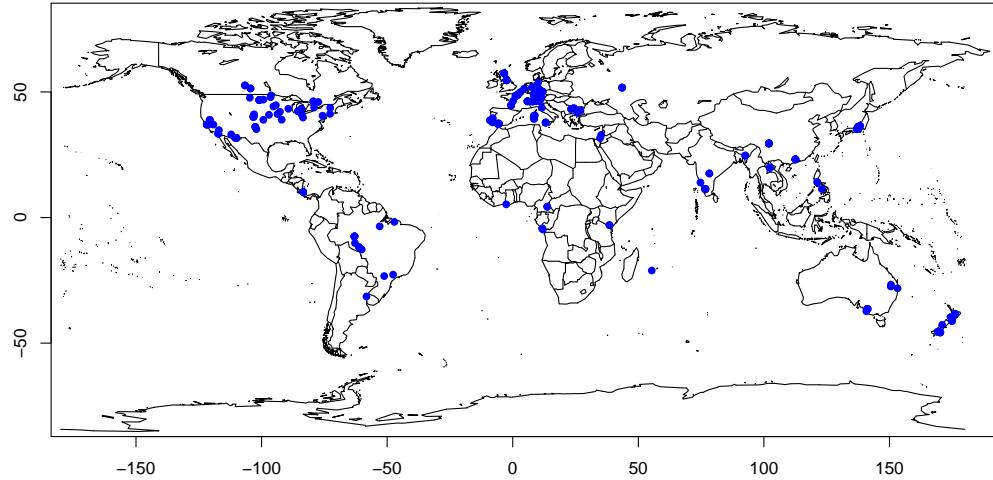


FIGURE 8 – Répartition géographique des sites échantillonés.

Pour chaque fosse, le sol est stratifié en plusieurs horizons dont on prélève simultanément un échantillon représentatif de chaque couche. Ainsi pour chaque profil, sont renseignés :

- la valeur du $\Delta^{14}\text{C}$ atmosphérique (en ‰) correspondant à l'année du prélevement.
 - les indications géographiques : latitude et longitude (en degré décimal), altitude (m).
 - les informations climatiques : précipitations moyennes annuelles (mm), précipitations moyennes de janvier et juillet (mm), températures annuelles moyennes ($^{\circ}\text{C}$), températures moyennes de janvier et juillet($^{\circ}\text{C}$), indice d'aridité.
- Les données climatiques sont issues du CRU (Climatic Research Unit) et croisées avec les données locales lorsque celles-ci sont disponibles dans l'article original.
- les descriptions du sol : stock de carbone en surface (kg.m^{-2}), stock de carbone du niveau échantillonné le plus profond (kg.m^{-2}).
 - le type de sol et la profondeur maximale théorique (cm) (dire d'experts, J.Balesdent).
 - le type d'écosystème associé au sol au moment de l'échantillonage.

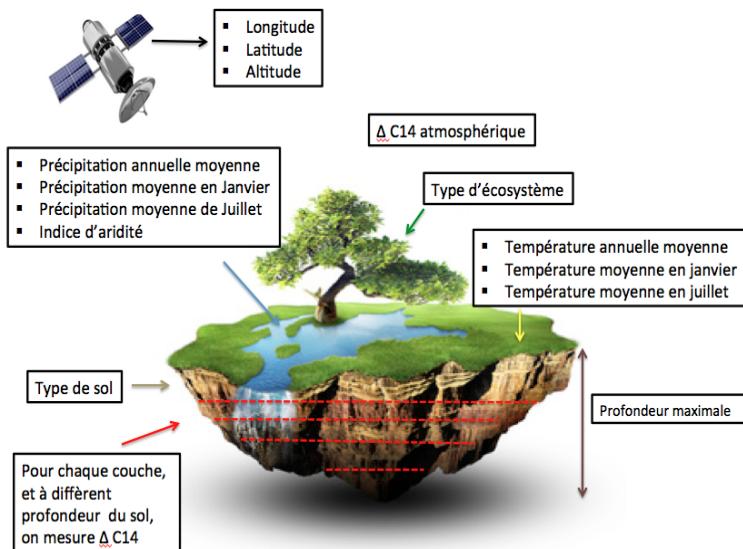


FIGURE 9 – Description de la base de données.

En ce qui concerne les types de sols (convention WRB) les profils sélectionnés pour l'estimation rassemblent 16 types différents : *andosol*, *arenosol*, *cambisol*,

chernozem, ferralsol, fluvisol, gleysol, kastanozem, leptosol, luvisol, nitisol, podzol et vertisol, phaeozem, acrisol, calcisol.

Les classes d'écosystème croisent deux caractéristiques physiques d'un site : la première fait référence au type d'écosystème rencontré à l'endroit du prélèvement (parmi : field, foret, grassland, savanna) et la seconde fait référence au type d'utilisation du sol (naturel vs cultivé).

On dispose ensuite des mesures de $\Delta^{14}C$ ainsi que des positions supérieures et inférieures de chacune des couches qui ont servi aux mesures (appelées "level.top" et "level.base"). Enfin, l'incertitude résultant de la mesure physique sur les valeurs de $\Delta^{14}C$ est parfois donnée. Voici la liste des noms des potentielles variables explicatives et leurs abréviations correspondantes (abréviations qui seront utilisées par la suite pour l'écriture des modèles) :

nom	abréviation	nom	abréviation
$\Delta^{14}C$	D14Catm	précipitation moyenne de janvier	Pjan
latitude	Lat	précipitation moyenne de juillet	PJuil
longitude	Long	indice d'aridité	Arid
altitude	Alt	profondeur maximale	Prof.max
précipitation annuelle moyenne	Pann	stock de carbone en profondeur	Stock.prof
température annuelle moyenne	Tann	stock de carbone en surface	Stock.surf
température moyenne de janvier	Tjan	gradient du stock de carbone	Pte.stock
température moyenne de juillet	Tjuil	type de sol	Sol
type d'écosystème	Land		

Traitement détaillé de la base de données Dans l'optique d'une estimation efficace, il a fallu éliminer certains profils, soit parce qu'ils ne contenaient que très peu de mesures de $\Delta^{14}C$ (moins de 3 mesures), soit parce que le traitement mécanique des couches ne permettait pas de prendre en compte les valeurs de $\Delta^{14}C$ obtenues (altérations dues à l'utilisation d'acide par exemple), soit enfin à cause des sols qui étaient de type "paléosol", ces derniers sont des sols anciens formés dans des conditions de climat et de végétation différentes de l'actuel et enterrés sous les dépôts épais plus récents, ils n'interviennent pas dans le cycle de carbone terrestre, raison pour laquelle on les élimine, on s'intéresse uniquement aux sols actifs. Enfin, après avoir défini une certaine couche du sol comme

valeur de référence pour l'origine de l'axe des profondeurs, les mesures réalisées au dessus de l'horizon O (litière) deviennent inutiles pour notre étude qui vise à décrire l'évolution en profondeur : il a donc fallu éliminer ces mesures très proches de la surface.

En résumé :

- éliminer les sites dont les valeurs de $\Delta^{14}C$ indiquées comme "modern".
- élimination des langues de gel (sites 86 et 87).
- élimination des triplicats (site 141-bulk).
- suppression des sites "paléosol", le travail actuel ne concerne que les sols actifs.
- acceptation des supports suivants : bulk, bulk after HCL, after concentrated HCL. Certaines études sont réalisées sur des molécules spécifiques ou des fractions granulométriques, densimétriques. . . non représentatives de la totalité de la matière organique du sol. Nous ne conservons que les mesures réalisées sur un support proche de la matière organique totale ("bulk").
- élimination des sites qui n'ont pas de données pour le $\Delta^{14}C$ (sites 144,145,149,150,186,187,188).
- élimination des sites de moins de trois mesures (sites 18, 72, 86, 144, 145, 149, 150, 162, 171, 178, 186, 187, 188, 198, 213, 218, 219, 234, 236, 239, 240, 242, 245, 246, 248, 249, 250, 251, 252, 253, 254, 255, 265, 287, 288, 289, 319, 340).

Parce que l'échantillonage de terrain n'est pas représentatif de l'extention réelle des sols et que la profondeur maximale des sols se révèle être un paramètre important, il faut réaliser l'analyse statistique sur un sous échantillon de la base de données, représentatif des sols réels.

Ainsi on considère tous les profils de plus de 100cm : ces profils correspondant aux points à droite de la ligne horizontale rouge en pointillées sur la figure suivante :

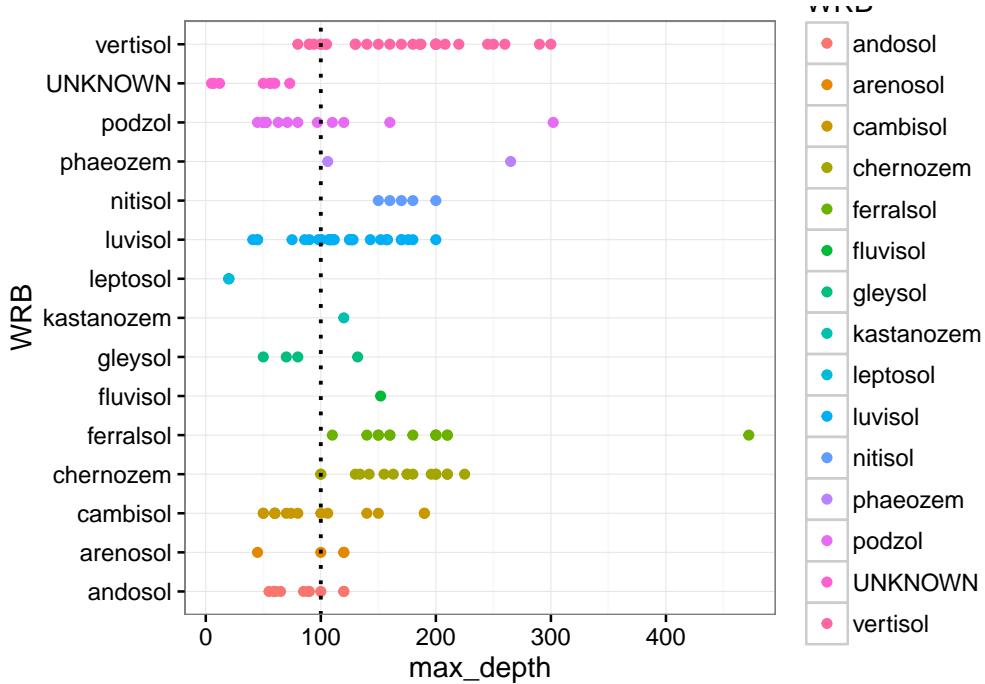


FIGURE 10 – Profondeur maximale des profils en fonction de types de sols (en abscisse), La courbe rouge en pointillé délimite les profils qui vont être sélectionnés pour l'estimation finale (ce sont ceux dont la mesure la plus profonde se situe suite à droite de la courbe rouge. Il en résulte une sélection de 159 profils.

Toutefois, ce seuil ne satisfait pas les caractéristiques des sols "courts". Aussi pour ceux-là, nous prenons les seuils suivants (dire d'experts, J.BALESDENT) :

1. Les leptosols- tous les sols.
2. Les podzols > 50cm.
3. les luvisols, gleysol, cambisol, andosol > 50cm.
4. le Kastenozem > 50cm.

On sélectionne ainsi 159 profils, voir annexe A.

Gestion des variables qualitatives La base de données regroupe 159 profils répartis en 15 types de sols différents et 12 combinaisons possibles pour la définition de l'écosystème (végétation + usage des terres). Certains types de sol sont sous-représentés et peuvent parfois être regroupés. Par ailleurs des associations

entre type de végétation et usage de sol ne sont pas possibles (ex. "field" et "natural"). Pour les profils restants après nettoyage de données, le nombre de profils par type est résumé dans les tableaux suivants :

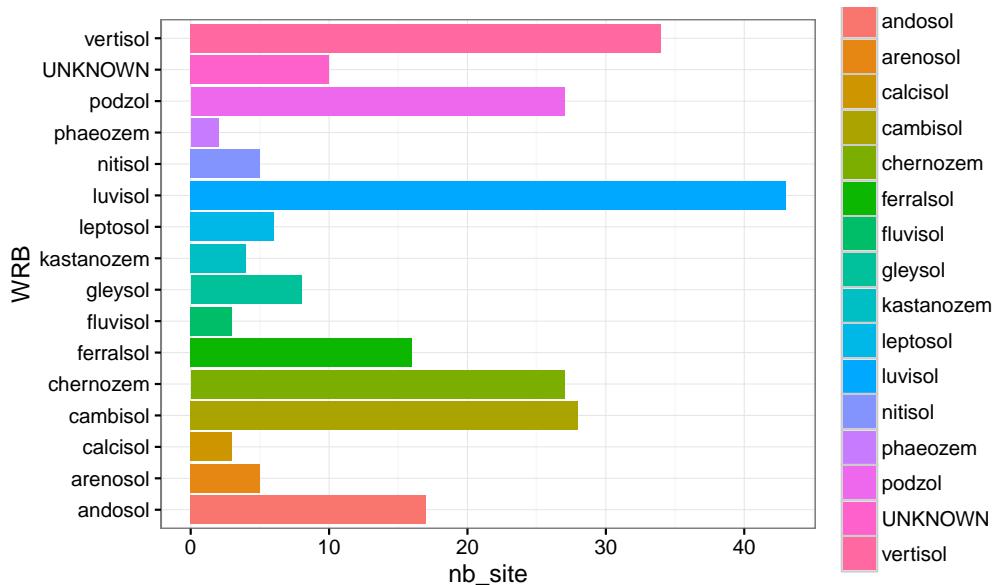


FIGURE 11 – Répartition du nbr de sites par type de sol.

	fluvisol	gleysol	kastanozem	leptosol	luvisol	andosol	arenosol
nb de profils	1	4	1	3	31	12	4
	cambisol	chernozem	ferralsol	phaeozem	calcisol	NA	
nb de profils	15	13	2	0	3	8	

TABLE 1 – Nombre de profils par type de sol, pour les 159 utilisés. Certaines catégories contiennent d'autres types de sols que celui indiqué en titre : "luvisol" contient les types "planosol" et "phaaeozem", et "ferralsol" contient le type "plinthosol".

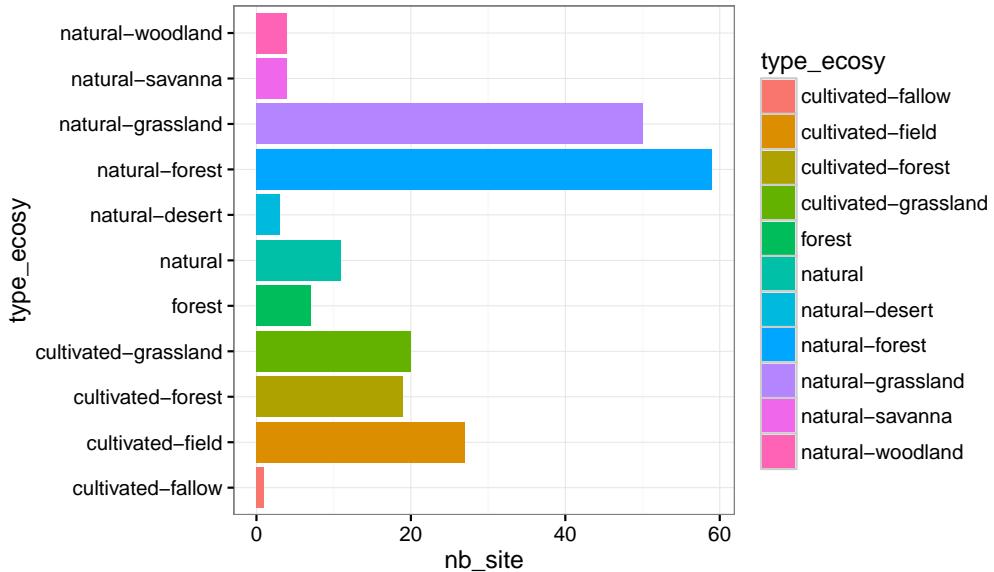


FIGURE 12 – Répartition du nbr de sites par type d'écosystème.

	grassland-cultivated	natural	savanna-natural	fallow-cultivated	field-cultivated
nb de profils	14	2	3	1	17
	forest	forest-cultivated	forest-natural	grassland-cultivated	NA
nb de profils	7	14	39	14	28

TABLE 2 – Nombre de profils par type d'écosystème, pour les 159 utilisés. NA représente les profils pour lesquels ni le type de végétation ni l'usage des terres ne sont renseignés.

3 Modélisation statistique des profils $\Delta^{14}C$

Afin de choisir un modèle qui décrit la dynamique de $\Delta^{14}C$ en fonction de la profondeur et en tenant en compte de tous les facteurs environnementaux, on peut tracer les profils de carbone des différents sites, voir la figure suivante :

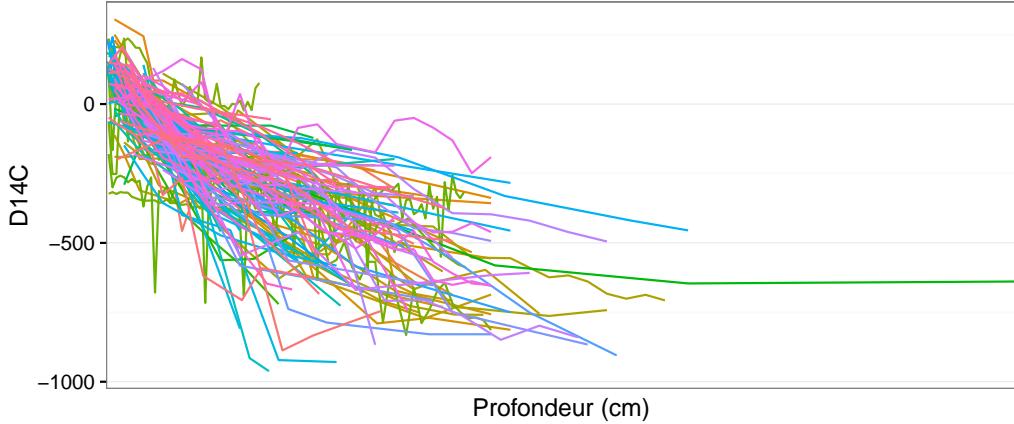


FIGURE 13 – Profils de carbone pour tous les sites étudiés.

On peut bien remarquer que la variabilité inter et intra site ainsi que les mesures réalisées sont différentes d'un site à un autre. Vu la non homogénéité de la variance entre les différents sites, on ne peut pas appliquer Anova afin de détecter l'influence de certains variables explicatives sur la dynamique de $\Delta^{14}C$.

Le premier travail consistait à reprendre le modèle proposé dans l'article de Jordane Mathieu et al. [2].

Dans cet article, le modèle choisi est le suivant : soit S le nombre total de sites. Pour un site $s \in [1, S]$, et pour une mesure $m \in [1, M_s]$ du site s , on modélise $\Delta^{14}C$ du sol en fonction de la profondeur :

$$\Delta^{14}C_m = \phi_{1,s} + \phi_{2,s} \exp\left(-\left(\frac{x_m}{\phi_{3,s}}\right)^{\phi_{4,s}}\right) + \sigma_s(x_m)\epsilon_m$$

avec $\epsilon_m \sim N(0, 1)$

- ϕ_1 : $\Delta^{14}C$ en grande profondeur.
- $\phi_1 + \phi_2$: $\Delta^{14}C$ en surface.
- ϕ_3 : lien avec la bosse à mi-hauteur.
- ϕ_4 : décroissance plus ou moins forte.

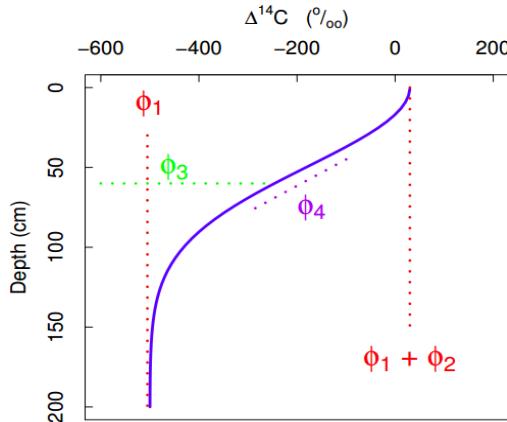


TABLE 3 – Représentation graphique du modèle choisi.

Ce modèle est suffisamment général pour intégrer diverses formes de courbes.

Du point de vu physique, la dynamique du carbone dans le sol peut être résumée en 3 processus mécaniques :

1. La diffusion : par définition c'est un phénomène de transport irréversible qui se traduit par la migration de la matière organique dans un milieu. Sous l'effet de l'agitation thermique, on observe un déplacement des constituants des zones de forte concentration vers celles de faible concentration.
2. L'advection : par définition se réfère surtout au transport vertical de la matière organique, dont le transport est dû au ruissellement d'eau.
3. La bioturbation : par définition désigne le phénomène de transfert de la matière organique par des êtres vivants au sein du sol.

Pour des raisons de difficulté de mesure, on ne profite pas de ces informations. D'où l'idée générale de considérer le modèle le plus adapté aux courbes de $\Delta^{14}\text{C}$ en exprimant ϕ_1, ϕ_2, ϕ_3 et ϕ_4 (variables latentes) en fonction des variables explicatives.

En appliquant ce modèle, on décrit d'une façon indirecte les processus mécaniques mentionnés ci-dessus. Mais comment ? En effet la diffusion dépend du type de sol et de type d'écosystème, la bioturbation dépend de la température alors que l'advection dépend des précipitations et de l'indice d'aridité.

4 Approches statistiques

4.1 Régression non-linéaire pour estimer les variables latentes-caractéristique de forme- $\phi_1, \phi_2, \phi_3, \phi_4$

Les variables latentes sont estimées via la fonction "optim" de R. Optim cherche à minimiser pour chaque site (s) et pour les différentes mesure $m \in [1, M_s]$, la fonction $f(s)$:

$$f(s) = \sum_{i=1}^m (\Delta^{14}C[s, m] - (\phi_{1,s} + \phi_{2,s} \exp(-(\frac{x_m}{\phi_{3,s}})^{\phi_{4,s}}))^2) \quad (1)$$

. Voici le choix des valeurs initiales pour l'algorithme :

1. $\phi_{1,0}$: La valeur de $\Delta^{14}C$ la plus profonde dont on dispose pour le site considéré.
2. $\phi_{2,0}$: La différence entre la valeur de $\Delta^{14}C$ en surface et celle en profondeur.
3. $\phi_{3,0}$: on utilise le lien entre ϕ_3 et la largeur de la demi-bosse à mi-hauteur : le maximum de la fonction est atteint en $x = 0$ et vaut $\phi_1 + \phi_2$ donc à mi-hauteur de la bosse, toujours en $x = 0$, il vaut $\phi_1 + \frac{\phi_2}{2}$. Ainsi la distance à mi-hauteur est égale à l'abscisse positive x_h qui assure la condition $\phi_1 + \phi_2 = \phi_1 + \phi_2 \exp(-\frac{x_h}{\phi_3})$. Cette abscisse vaut $x_h = \phi_3 \ln(2)^{\frac{1}{\phi_4}}$. Il suffit alors de connaître approximativement x_h à partir des données pour pouvoir exprimer $\phi_{3,0}$ par : $\phi_{3,0} = x_h \ln(2)^{\frac{-1}{\phi_4,0}}$.
4. ϕ_4 semble prendre ses valeurs entre 1 et 3.

Le code R et le tableaux de valeurs estimées sont donnés en annexe B.

Le graphique suivant illustre la distribution des valeurs des variables latentes via optim :

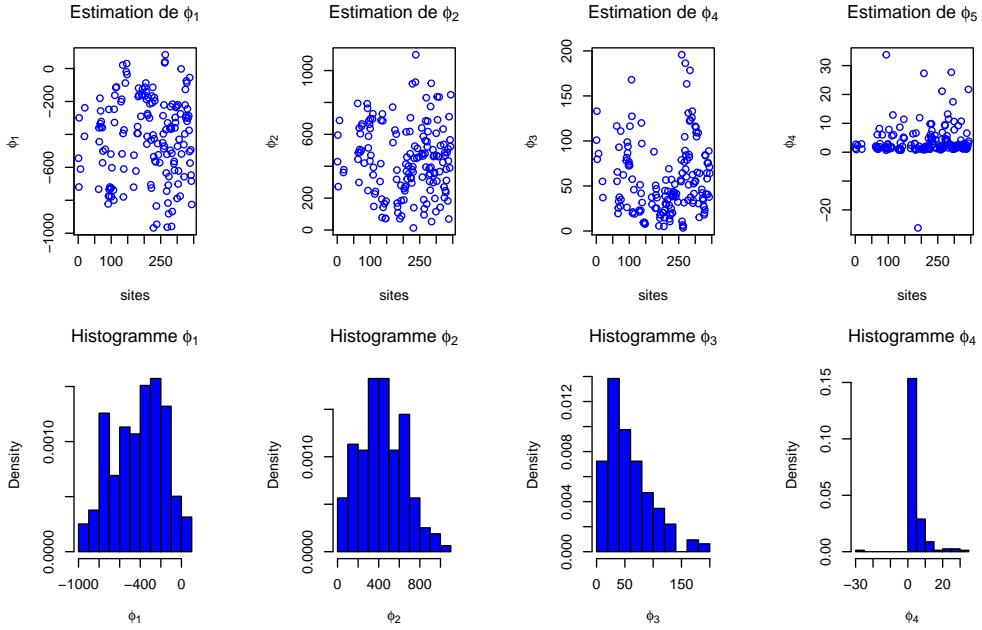


FIGURE 14 – Valeurs et histogarammes des variables- ou caractéristiques de forme - estimées via "optim" ϕ_1, ϕ_2, ϕ_3 et ϕ_4 pour l'ensemble des 159 sites de l'étude.

5 Variabilité intra-site

Le modèle non linéaire normal à variance hétérogène d'occurrence locale des mesures s'écrit de la façon suivante :

Pour un site $s \in [1 : S]$, et pour une mesure $m_s \in [1 : n_s]$, on modélise l'évolution de $y = \Delta^{14}C$ du sol en fonction de la profondeur $x(m_s)$ par :

$$y(m_s) = \varphi_1(s) + \varphi_2(s) \exp - \left(\frac{x(m_s)}{\varphi_3(s)} \right)^{\varphi_4(s)} + \sigma(m_s) \times \varepsilon(m_s) \quad (2)$$

$$\varepsilon(m_s) \sim N(0, 1)$$

On voit que ce modèle possède la même structure pour chaque site, mais il est non linéaire car la réponse y est une fonction non linéaire des coefficients $\varphi_1, \varphi_2, \varphi_3, \varphi_4$. C'est normal puisque le bruit de mesure ε est supposé normal. Dans un premier temps on posera une variance homogène $\sigma(m_s) = \sigma$. Pour des raisons de positivité, j'utilise par la suite des variables caractéristiques transfor-

mées :

$$\begin{aligned}\theta_1(s) &= \varphi_1(s) \\ \theta_2(s) &= \varphi_2(s) \\ \theta_3(s) &= \log \varphi_3(s) \\ \theta_4(s) &= \log \varphi_4(s)\end{aligned}$$

Dans le langage des physiciens, ces grandeurs caractéristiques $\theta_1, \theta_2, \theta_3, \theta_4$ s'appellent improprement des paramètres. En fait ce sont des variables latentes, elles sont tirées dans une même loi de probabilité, ici une loi normale de dimension 4, dont la variance régit la variabilité entre sites.

5.1 Variabilité intersite

$\theta_1, \theta_2, \theta_3, \theta_4$ sont des variables intermédiaires imaginées par le modélisateur qui règlent le lien entre des variables explicatives et des réplications (faîtes sur divers sites) d'un même type de mesures. On crée donc un modèle de covariation des 4 caractéristiques $\theta(s) = (\theta_1(s), \theta_2(s), \theta_3(s), \theta_4(s))$ du comportement vis à vis du $\Delta^{14}C$ d'un site s . Il faut d'abord réajuster chaque caractéristique de telle sorte que chaque site puisse être considéré comme une réplication aléatoire d'un de ses voisins : pour cela, on effectue ici la soustraction d'une fonction linéaire (régression) de P proxys enregistrés pour le site, variables explicatives quantitatives et qualitatives encodés dans F , matrice de S lignes et P colonnes. Ce modèle s'écrit de la façon suivante :

Pour chaque caractéristique $j \in [1 : 4]$, on suppose une réponse de type normale multivariée

$$\theta_j(s) - \sum_{p=1}^P F(s, p) \beta(p, j) = E_j(s) \quad (3)$$

On peut bien sûr mettre le coefficient $\beta(p, j)$ à 0 si la $j^{\text{ème}}$ caractéristique n'a aucune raison *a priori* de dépendre du $p^{\text{ème}}$ proxy. L'effet aléatoire $E(s) =$

$\begin{pmatrix} E_1(s) \\ E_2(s) \\ E_3(s) \\ E_4(s) \end{pmatrix}$ manifeste la partie explicative non prise en compte par les proxys. Il

s'agit d'une variabilité provenant de la plus ou moins grande ressemblance entre les tirages, réglée par la variance de la loi des E . Pour rester dans le domaine normal, on suppose que ces effets sont tirés selon une loi multinormale de matrice

4×4 de variance covariance Ω .

$$E(s) \sim N_4(0, \Omega)$$

Du point de vue du langage de la statistique, seuls σ , Ω et β sont les paramètres du modèle formé par les équations 3+2.

6 Lecture et Nettoyage des données

On a les informations collectées pour 159 sites (voir annexe A). Pour l'estimation on enlève les sites suivants : 190 (ϕ_4 estimé est négatif), 146, 147, 148 et 238. Ainsi les lignes de tableaux correspondent aux valeurs manquantes du $\Delta^{14}C$ et des principales variables explicatives. En plus, je supprime tous les sites ayant comme type de sol "UNKNOWN" ou comme type d'écosystème NA. Après sélection, il nous reste 104 sites, les sites éliminés sont représentés en annexe.

Préparation des grandeurs explicatives En résumé, les 8 variables explicatives quantitatives prises en compte dans notre modèle sont : la valeur atmosphérique de l'année de prélèvement, la latitude, l'indice d'aridité, le stock de carbone en surface et en profondeur, la température et la précipitation annuelle moyenne du site , ainsi que la différence en valeur absolue de la température entre janvier et juillet.

Il nous reste 2 variables catégorielles : le type de sol et le type d'écosystème. Vu le nombre de profils par type de sol et par type d'écosystème, on a regroupé les 3 types de sol suivant "fluvisol(1 profil)", "kastanozem (1 profil)", "phaeozem(2 profils)" en un seul groupe. De même pour les 2 types d'écosystème "natural-desert(1 profil)" et "natural(2 profils)" qui sont également regroupés en un seul groupe. Par contre, le type de sol "leptosol" n'est pas représenté dans cette étude. On a 3 profils de $\Delta^{14}C$ pour le type "leptosol" et ils sont considérés comme courts vu la profondeur maximale mesurée. Le code R est en annexe C

7 Inférence bayésienne sous Jags

Afin d'estimer nos variables latentes on considère le modèle bayésien hiérarchique suivant, en tenant compte des incertitudes intra et inter site :

1. Modèle :

$$y_{[s,z]} \sim N(f(\phi_s, s, z), \sigma^2); \quad s \in [1, S] \text{ et } z \in [1, m_s] \quad (4)$$

avec $y_{[s,z]}$: mesure $\Delta^{14}C$ pour le site s au profondeur z , $f(\phi_s, s, z)$ le modèle ajusté au site s à la profondeur z . On suppose que l'erreur de mesure σ^2 est homogène pour tous les sites.

2. Variables latentes : on a pour chaque site :

$$\phi_s \sim MN(X[s,] * \Theta_{[26,4]}, \Omega) \quad (5)$$

3. Prior :

$$\begin{aligned} \tilde{\sigma}^2 &\sim Gamma(a, b) \\ i \in [1, 26] \text{ et } j \in [1, 4] \Theta[i, j] &\sim N(\mu, \tau^2) \\ \Omega &\sim Wishart(v, R_{[4,4]}) \end{aligned}$$

Voici le graphe DAG, ainsi que le modèle utilisé sous Jags (sauvegardé sous fichier.txt) : notez que la commande dflat() n'existe pas sous Jags, elle peut être substituée par une loi normale en supposant une variance de valeur importante.

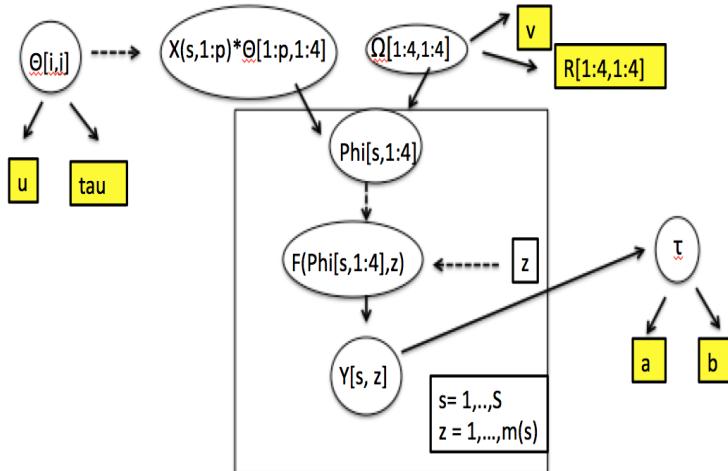


FIGURE 15 – Les quantités aléatoires sont entourées par des ellipses et les quantités fixes ou observées sont entourées par des rectangles.

Le modèle utilisé sous Jags

```
#####
model{

# variables latentes

for (i in 1:N){
  phi[i,1:4] ~ dmmnorm(mu[i,],Prec[,])
  mu[i,1] <- inprod(X[i,1:p] ,theta[1:p,1])
  mu[i,2] <- inprod(X[i,1:p] ,theta[1:p,2])
  mu[i,3] <- inprod(X[i,1:p] ,theta[1:p,3])
  mu[i,4] <- inprod(X[i,1:p] ,theta[1:p,4])
}

# Vraisemblance

for (i in 1:N){
  for (j in 1:size_par_site[i]){
    y[i,j] ~ dnorm(m[i,j],tau)
    m[i,j] <- phi[i,1]+phi[i,2]*exp(-
      pow((z[i,j]/exp(phi[i,3])),exp(phi[i,4])))
  }
}

# Prior

for (j in 1:4){
  for (i in 1:p){
    theta[i,j] ~ dnorm(0,0.0001)  #loi impropre
  }
}

Prec[1:4,1:4] ~ dwish(R[,],v)

tau ~ dgamma(a,b)
}
#####
```

Pour faire tourner le modèle sous Jags on a besoin d'attribuer des valeurs initiales pour : Θ , Ω et σ^2 . Afin d'accélérer la convergence, il est préférable de donner des valeurs initiales qui sont vraisemblables, ainsi on a réalisé une analyse de variance multivariée (Manova) pour initialiser la matrice $\Theta_{[p,4]}$ où p est le nombre de variables explicatives avec des valeurs intéressantes. Comme variable à expliquer, on a utilisé la matrice Φ (ncol = p et nrow = 104) où la ligne i de cette matrice correspond aux caractéristiques du modèle d'ajustement du site i (obtenu par optim).

8 Tourner le modèle sur des données simulées

Une fois Θ et Ω sont initialisés, on peut simuler les variables latentes (ϕ_1, ϕ_2, ϕ_3 et ϕ_4) pour chaque site s en tirant dans une loi multinormale suivant l'équation 5.

Une fois les caractéristiques de chacun de ces sites sont tirées, on peut simuler nos données de $\Delta^{14}C$ en tirant dans une loi normale suivant l'équation 4. Le code R est donné an annexe D.

Suite aux simulations, on peut avoir recours à Jags qui nous permet de calculer la loi *a posteriori* des paramètres à estimer en utilisant les algorithmes de Metropolis Hastings. Le code R est donné en annexe E.

Choix des priors : Concernant les hyperparamètres de la loi gamma de la précision $\tilde{\sigma}^2$, on a choisi une loi gamma non-informative avec $a = 0.001$ et $b = 0.001$. Pour chaque élément de la matrice Θ , on a supposé une loi normale vague car à priori on a aucune information. Il reste les hyperparamètres de la loi de Wishart de la matrice de précision, on a supposé un degré de liberté de 5 pour que la matrice de précision soit proche du paramètre d'échelle (une matrice de taille 4*4). Le paramètre d'échelle est la matrice de précision obtenue par Manova.

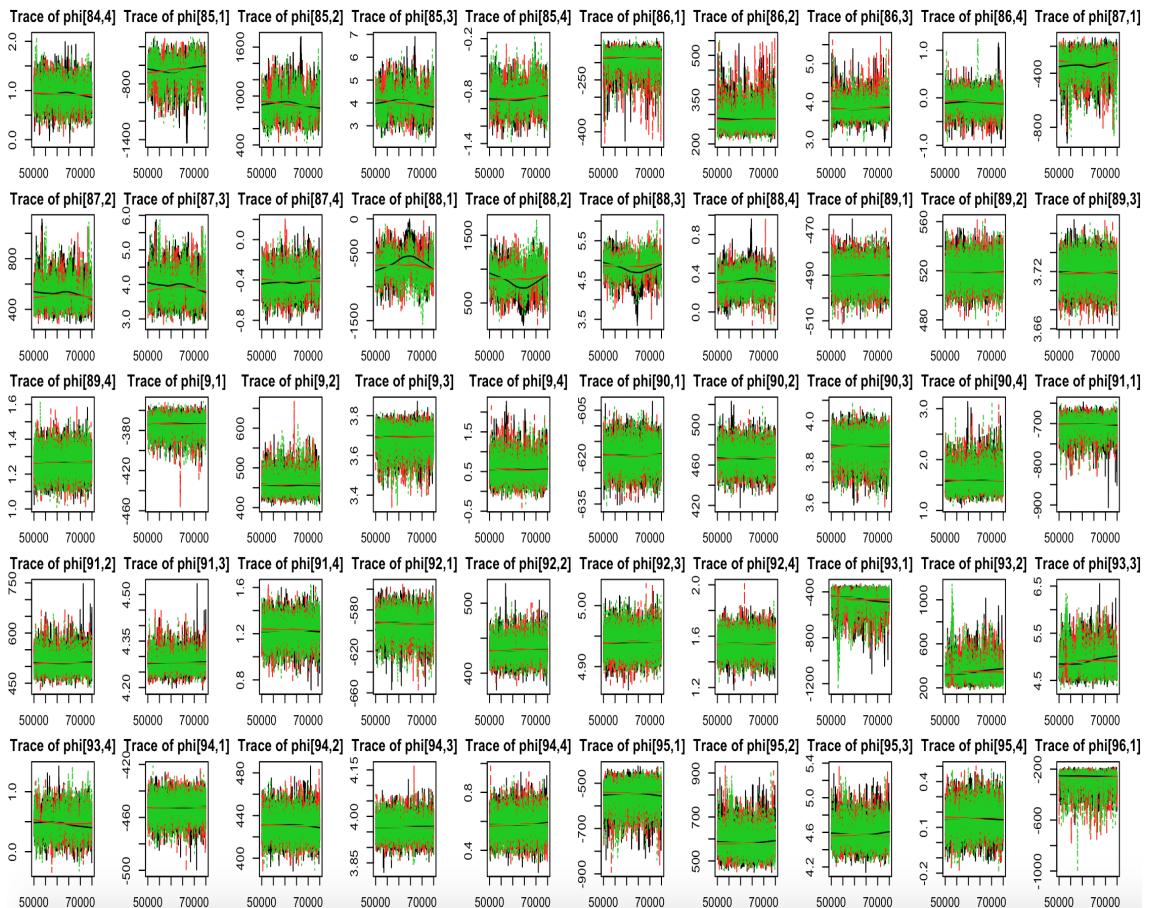
8.1 Diagnostic de la convergance sur les données simulées

Mixing Dans la théorie des chaînes de Markov, on s'attend à ce que nos 3 chaînes MCMC convergent en théorie vers la distribution stationnaire, qui est aussi notre distribution cible. Cependant, il n'y a aucune garantie pratique que notre chaîne converge après M itérations. Comment peut on savoir si notre chaîne a effectivement convergé ? On ne peut jamais être sûr, mais il y a plusieurs tests que nous pouvons faire, à la fois visuels et statistiques, pour voir si

la chaîne semble avoir convergé.

Tous les diagnostics se trouvent dans le package coda de R. Mais avant il faut transformer nos 3 chaines en "objets MCMC".

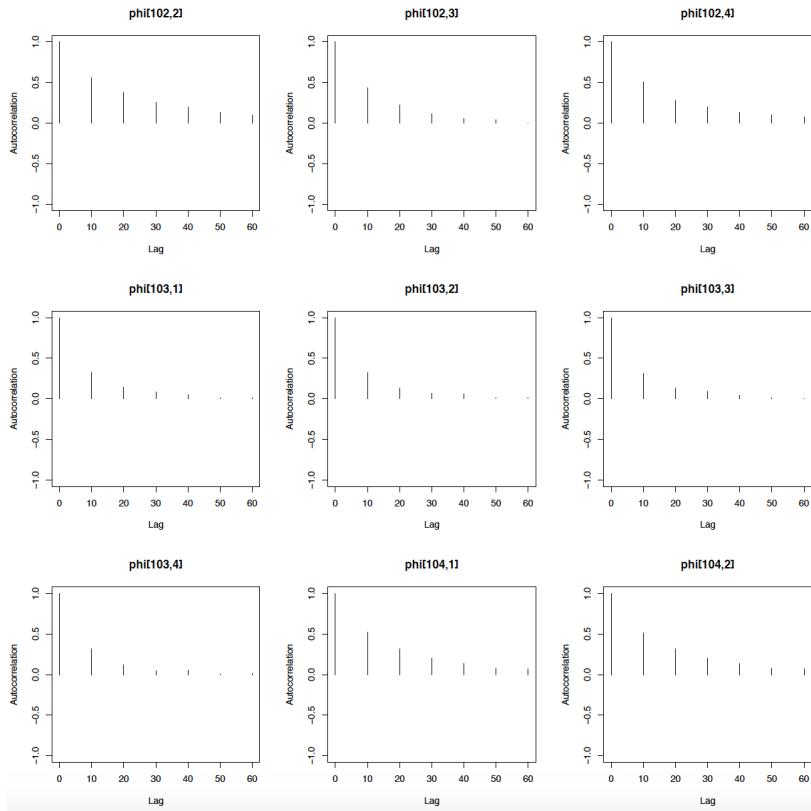
Une façon de voir si notre chaîne a convergé est de voir comment les 3 chaînes sont mélangées (mixing) ou comment elles se déplacent autour de l'espace des paramètres. Si notre chaîne prend beaucoup de temps pour se déplacer dans l'espace des paramètres, il faudra plus de temps à converger. Un traceplot est une représentation graphique de la valeur du paramètre à chaque itération. Nous pouvons voir si notre chaîne est coincée dans certaines zones de l'espace des paramètres, ce qui indique un mauvais mélange. Voir la figure ci-dessous :



Autocorrélation En plus on peut évaluer la convergence en terme des corrélations entre les tirages de notre chaîne de Markov. L'autocorrélation ρ_k : est la corrélation entre chaque tirage et celui correspondant à un décalage de K. Voir l'équation suivante :

$$\rho_k = \frac{\sum(x_i - \bar{x})(x_{i+k})}{\sum(x_i - \bar{x})^2} \quad (6)$$

On s'attend à ce que l'autocorrélation ρ_k soit plus petite lorsque K augmente. Si l'autocorrélation est encore relativement élevée, pour des valeurs plus élevées que k, on a un degré haut de corrélation et un mélange lent. Voir la figure suivante :



Gelman et Rubin Le test de Gelman et Rubin nous permet de calculer la variance intra (W) et inter (B) chaines MCMC, ainsi on définit la variance de la

distribution stationnaire par :

$$\tilde{Var}(\theta) = \left(1 - \frac{1}{n}\right)W + \frac{1}{n}B \quad (7)$$

Et le potential scale reduction factor \tilde{R} est défini par :

$$\tilde{R} = \sqrt{\frac{\tilde{Var}(\theta)}{n}} \quad (8)$$

Si \tilde{R} est élevé (supérieur à 1.1 ou 1.2), nous devrions exécuter nos chaînes plus longtemps pour améliorer la convergence. Dans mon cas tous les \tilde{R} sont inférieurs à 1.2. Le code R est donnée en annexe F.

8.2 Comparaison entre la matrice Θ_{true} et $\Theta_{posterior}$

Pour vérifier l'algorithme tourne bien, il faut que la valeur de la matrice Θ à postériori ne soit pas loin de la vraie matrice Θ à partir de laquelle on a simulé les données. Le code R et les graphes sont donnés en annexe G.

9 Application finale sur les données réelles de $\Delta^{14}C$

On profite des données réelles de 104 profils de $\Delta^{14}C$. 74 sites ont servi pour l'apprentissage et 30 pour la validation. Les bandeaux de prédictions qui illustrent les résultats sur l'ensemble d'apprentissage (74 sites) ainsi que l'ensemble de validation (30 sites) sont donnés en annexe H (en orange, résultant de notre connaissance uniquement partielle des paramètres) auxquels se superpose une incertitude sur la mesure des profils (ici supposée de variance homogène) représentée en grise.

10 Sélection du modèle dans le cadre bayésien

Un problème crucial dans la construction d'un modèle de régression multiple est la sélection des prédicteurs à intégrer dans le modèle. Plus précisément, étant donné une variable dépendante Y et un ensemble de prédicteurs potentiels, le

problème consiste à identifier le meilleur modèle de forme, par exemple dans notre cas :

$$\phi_1 = X_1^* \beta_1^* + \cdots + X_q^* \beta_q^* \quad (9)$$

Où X_1^*, \dots, X_q^* est le sous ensemble sélectionné parmi X_1, \dots, X_p . Plusieurs critères de sélection basés sur la comparaison de tous les 2^p sous-modèles possibles sont considérés comme AIC, Cp et BIC.

Maleureusement, si p est grand les exigences de calcul pour ces procédures peuvent être prohibitives.

Dans cette partie, on introduit la procédure SSVS (Stochastic Search Variable Selection). SSVS consiste à intégrer la régression entière dans un modèle hiérarchique bayésien de mélange normal, où les variables latentes sont utilisées pour identifier le sous ensemble choisi. Ainsi le meilleur sous ensemble de prédicteurs est composé de ceux ayant une probabilité *a posteriori* importante. Cette technique est basée sur l'échantillonnage de Gibbs afin de générer indirectement des répliques de la loi multinomiale *a posteriori* sur l'ensemble des choix possibles. Le sous ensemble de prédicteurs avec la plus élevée probabilité peut être identifié par son apparition fréquente dans l'échantillonneur de Gibbs.

Un modèle hiérarchique de sélection de variables : Pour un modèle de régression, on considère que la variable à expliquer Y suit la loi suivante :

$$Y|\beta, \sigma^2 \sim N(X\beta, \sigma^2 I) \quad (10)$$

où Y est $n \times 1$, $X = [X_1, \dots, X_p]$ est $n \times p$, $\beta = (\beta_1, \dots, \beta_p)$ et σ^2 est un scalaire. β et σ^2 sont considérés comme inconnus. Pour le modèle considéré, sélectionner un sous ensemble de prédicteurs est équivalent à affecter des zéros aux β_i correspondants aux prédicteurs non sélectionnés.

Ainsi, on considère le modèle précédent comme une partie d'un large modèle hiérarchique où on suppose que chaque composante de β provient d'un modèle de mélange de 2 distributions normales de variances différentes en introduisant une variable latente $\gamma_i = 0$ ou 1 , on représente le modèle de mélange normal par :

$$\beta_i | \gamma_i \sim (1 - \gamma_i)N(0, \tau_l) + \gamma_i N(0, c_i^2 \tau_i^2) \quad (11)$$

et

$$P(\gamma_i = 1) = 1 - P(\gamma_i = 0) = p_i \quad (12)$$

Lorsque $\gamma_i = 0$, $\beta_i \sim N(0, \tau_l)$, et lorsque $\gamma_i = 1$, $\beta_i \sim N(0, c_i^2 \tau_i^2)$. L'interprétation de cette formule se fait de la manière suivante :

Dans un premier temps on suppose que τ_l est très petite, ainsi si $\gamma_i = 0$, β_i va être probablement très petit et peut être estimé par 0. Dans un second temps, on suppose que c_i est assez large, ainsi si $\gamma_i = 1$, β_i est loin d'être estimé par 0. p_i décrit notre croyance *a priori* qui suppose que β_i est un estimateur non nul, ce qui est équivalent à dire que X_i doit être inclus dans le modèle.

Pour obtenir (11) comme un prior de $\beta_i|\gamma_i$ on suppose une loi multinormale *a priori*

$$\beta|\gamma \sim N_p(0, D_\gamma R D_\gamma) \quad (13)$$

où $\gamma = (\gamma_1, \dots, \gamma_p)$, R est la matrice de corrélation à priori et

$$D_\gamma = \text{diag}[a_1\tau_1, \dots, a_p\tau_p] \quad (14)$$

avec $a_i = 1$ si $\gamma_i = 0$ et $a_i = c_i$ si $\gamma_i = 1$. Pour γ on a supposé un modèle de Bernouilli avec une probabilité p_i (la probabilité que β_i soit non nul).

Finalement on suppose une loi inverse gamma pour modéliser la précision $\tilde{\sigma}^2$:

$$\tilde{\sigma}^2 \sim G\left(\frac{v}{2}, \frac{v\lambda}{2}\right) \quad (15)$$

Identifier le meilleur modèle avec $f(\gamma|Y)$ L'intérêt principal de ce modèle hiérarchique bayésien est de chercher la loi marginale *a posteriori*

$$f(\gamma|Y) \propto f(Y|\gamma)f(\gamma) \quad (16)$$

Un choix raisonnable est de considérer que l'inclusion du prédicteur X_i est indépendant, de l'inclusion du prédicteur X_j , pour tout $i \neq j$ ainsi on peut écrire :

$$f(\gamma) = \prod p_i^{\gamma_i} (1 - p_i)^{(1-\gamma_i)} \quad (17)$$

Un prior uniforme est un cas spécial de (17) où $f(\gamma) = \frac{1}{2^p}$, ainsi la probabilité d'inclusion est : $p_i = \frac{1}{2}$. Le travail précédent était représenté dans l'article (George and McCulloch 1993) [1].

Après cette vision globale de la technique de sélection de variable revenant à mon cas d'étude.

On a choisi $c = 1000$ (assez large) et $R = I$. A noter que les variables à expliquer sont les variables latentes ϕ_1, ϕ_2, ϕ_3 et ϕ_4 . Le modèle sous Jags est le suivant :

```

model{
  # typical regression priors
  sd_y ~ dunif(0, 100)
  tau_y <- pow(sd_y, -2)
  Prec[1:4,1:4] ~ dwish(R[,],5)

  # SSVS technique

  # Prior variance beta

  sd_bet ~ dunif(0, 100)
  tau_in[1] <- pow(sd_bet, -2)    # coef effectively zero
  tau_in[2] <- tau_in[1]/ 1000 # nonzero coef

  # prior probabilité d'inclusion

  p_ind[1] <- 1/2
  p_ind[2] <- 1 - p_ind[1]

  for (i in 1:ncov){
    ind1[i] ~ dcat(p_ind[]) # returns 1 or 2
    gamma1[i] <- ind1[pos[i]] - 1    # returns 0 or 1

    ind2[i] ~ dcat(p_ind[]) # returns 1 or 2
    gamma2[i] <- ind2[pos[i]] - 1    # returns 0 or 1

    ind3[i] ~ dcat(p_ind[]) # returns 1 or 2
    gamma3[i] <- ind3[pos[i]] - 1    # returns 0 or 1

    ind4[i] ~ dcat(p_ind[]) # returns 1 or 2
    gamma4[i] <- ind4[pos[i]] - 1    # returns 0 or 1

  # Prior beta
}

```

```

beta[i,1] ~ dnorm(0, tau_in[ind1[i]])
beta[i,2] ~ dnorm(0, tau_in[ind2[i]])
beta[i,3] ~ dnorm(0, tau_in[ind3[i]])
beta[i,4] ~ dnorm(0, tau_in[ind4[i]])
}

# Latent variable

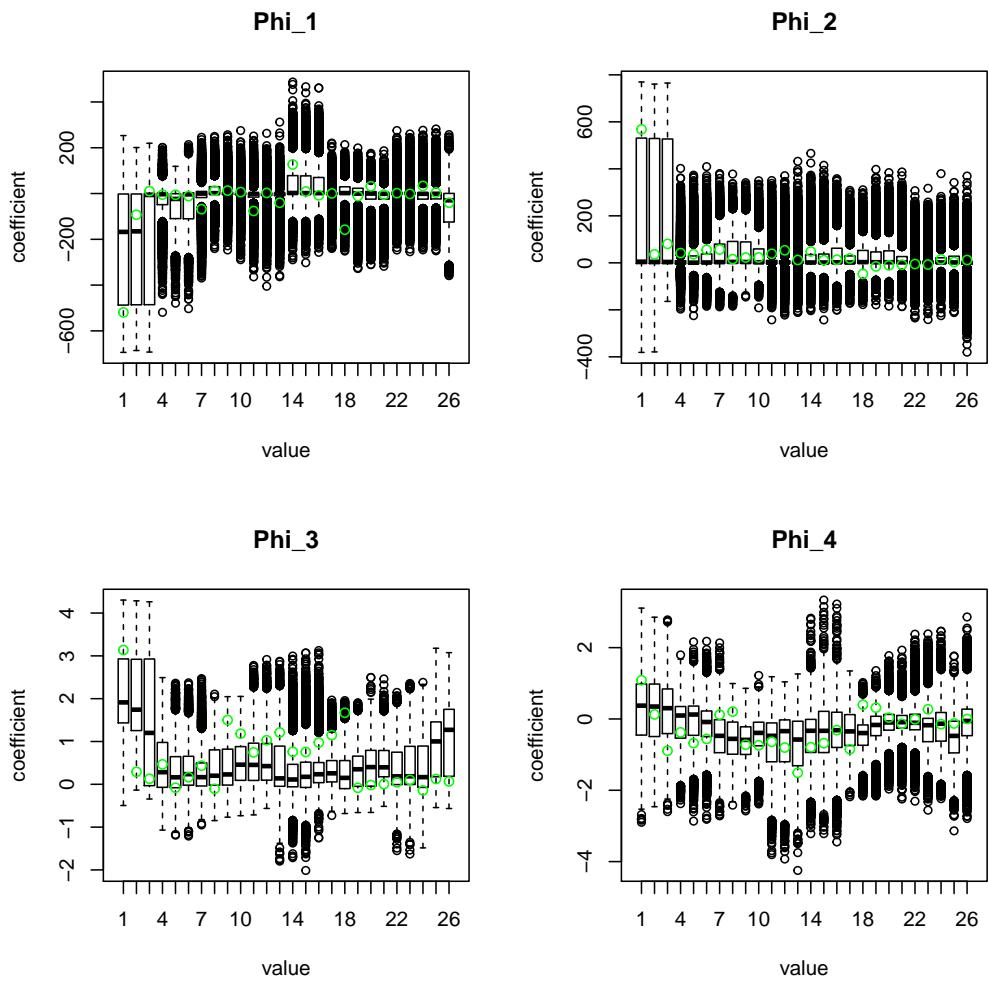
for (i in 1:N){
  phi[i,1:4] ~ dmmnorm(mu[i,],Prec[,])
  mu[i,1] <- inprod(X[i,] ,beta[,1])
  mu[i,2] <- inprod(X[i,] ,beta[,2])
  mu[i,3] <- inprod(X[i,] ,beta[,3])
  mu[i,4] <- inprod(X[i,] ,beta[,4])
}

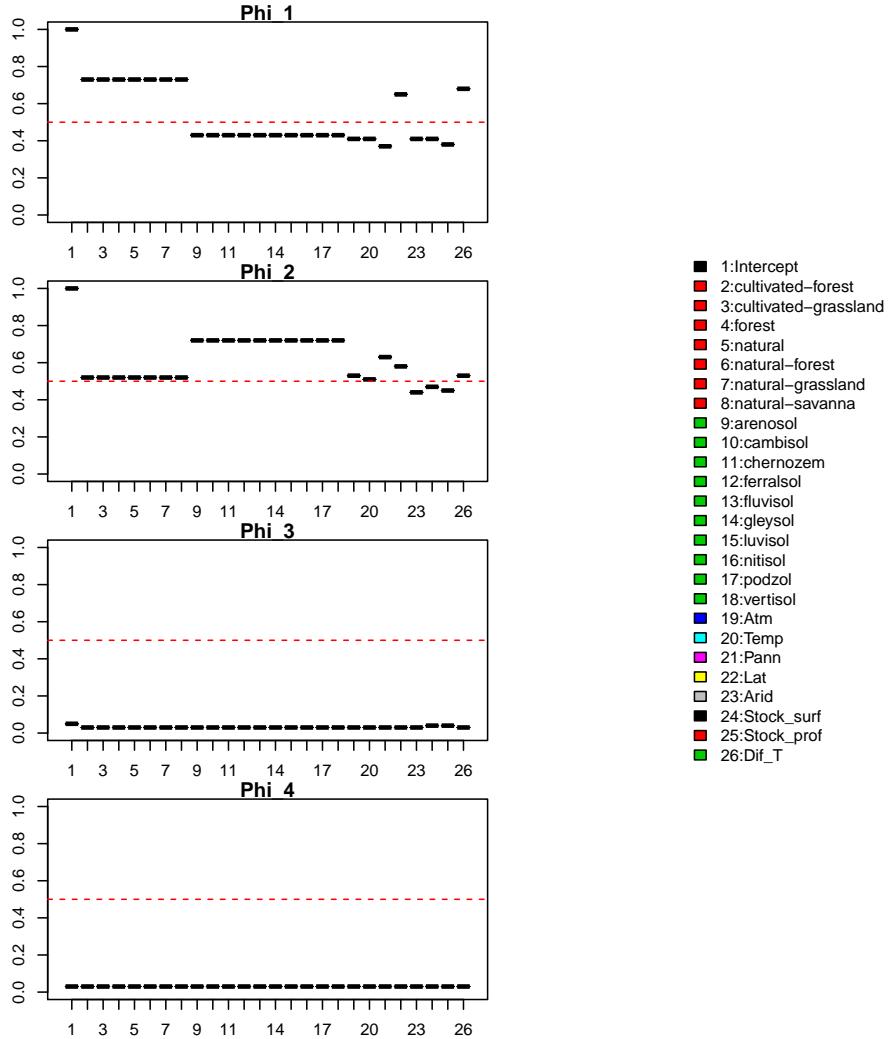
# likelihood

for (i in 1:N){
  for (j in 1:size_par_site[i]){
    y[i,j] ~ dnorm(m[i,j],tau_y)
    m[i,j] <- phi[i,1]+phi[i,2]*exp(-
      pow((z[i,j]/exp(phi[i,3])),exp(phi[i,4])))
  }
}

```

Les graphes ci-dessous nous permet de voir l'influence de chacune des variables explicatives (qualitatives et quantitatives) sur nos variables latentes, le code R est donné en annexe I :





ϕ_1 est interprétée comme étant la concentration de $\Delta^{14}\text{C}$ en profondeur, ainsi on peut dire que les facteurs qui influent le plus sur cette concentration sont : en premier lieu le type d'écosystème (avec une probabilité de 0.75 d'être sélectionné dans le modèle), suivi de la latitude (0.65) et la différence de température entre les mois de janvier et juillet (0.68). Le type de sol contribue de 43% à la compréhension de la dynamique du carbone en profondeur. Voir le tableau ci-dessous pour les autres covariables. Pour ϕ_2 qui représente la concentration de $\Delta^{14}\text{C}$ à la surface, c'est le type de sol qui impacte le plus cette concentration (avec une probabilité *a posteriori* 0.72 d'être sélectionné dans le modèle), suivi par le taux de précipitation annuel (0.62), la latitude (0.58), la différence de température entre

juillet et janvier (0.53), la concentration atmosphérique en $\Delta^{14}C$ (0.53) et finalement la température annuelle avec un eprobabilité de 0.51.

Pour le reste voir le tableau ci-dessous. Pour les deux variables latentes restantes ϕ_3 et ϕ_4 , aucun parmis les prédicteurs potentiels n'est significatif.

	Covariables	Phi_1	Phi_2	Phi_3	Phi_4
1	Intercept	1.00	1.00	0.05	0.03
2	cultivated-forest	0.73	0.52	0.03	0.03
3	cultivated-grassland	0.73	0.52	0.03	0.03
4	forest	0.73	0.52	0.03	0.03
5	natural	0.73	0.52	0.03	0.03
6	natural-forest	0.73	0.52	0.03	0.03
7	natural-grassland	0.73	0.52	0.03	0.03
8	natural-savanna	0.73	0.52	0.03	0.03
9	arenosol	0.43	0.72	0.03	0.03
10	cambisol	0.43	0.72	0.03	0.03
11	chernozem	0.43	0.72	0.03	0.03
12	ferralsol	0.43	0.72	0.03	0.03
13	fluvisol	0.43	0.72	0.03	0.03
14	gleysol	0.43	0.72	0.03	0.03
15	15:luvisol	0.43	0.72	0.03	0.03
16	16:nitisol	0.43	0.72	0.03	0.03
17	podzol	0.43	0.72	0.03	0.03
18	vertisol	0.43	0.72	0.03	0.03
19	Atm	0.41	0.53	0.03	0.03
20	Temp	0.41	0.51	0.03	0.03
21	Pann	0.37	0.63	0.03	0.03
22	Lat	0.65	0.58	0.03	0.03
23	Arid	0.41	0.44	0.03	0.03
24	Stock_surf	0.41	0.47	0.04	0.03
25	Stock_prof	0.38	0.45	0.04	0.03
26	Dif_T	0.68	0.53	0.03	0.03

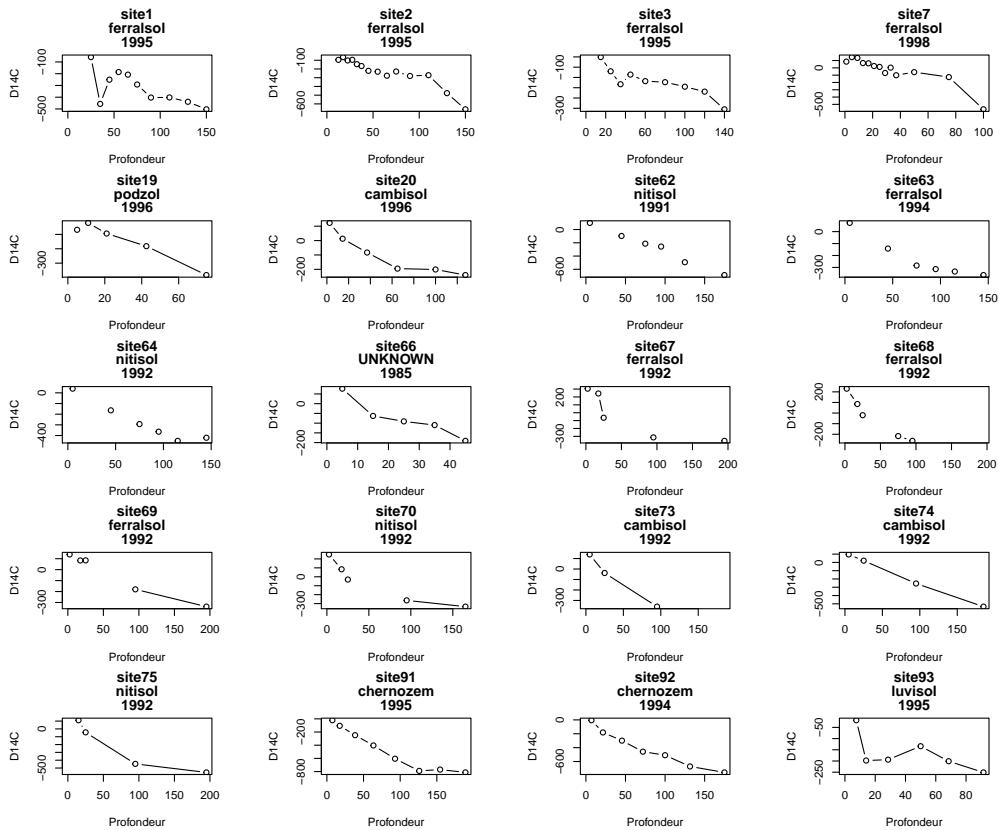
FIGURE 16 – Probabilité de sélection *a posteriori* dans le modèle pour chacune des variables latentes - caractéristiques du modèle.

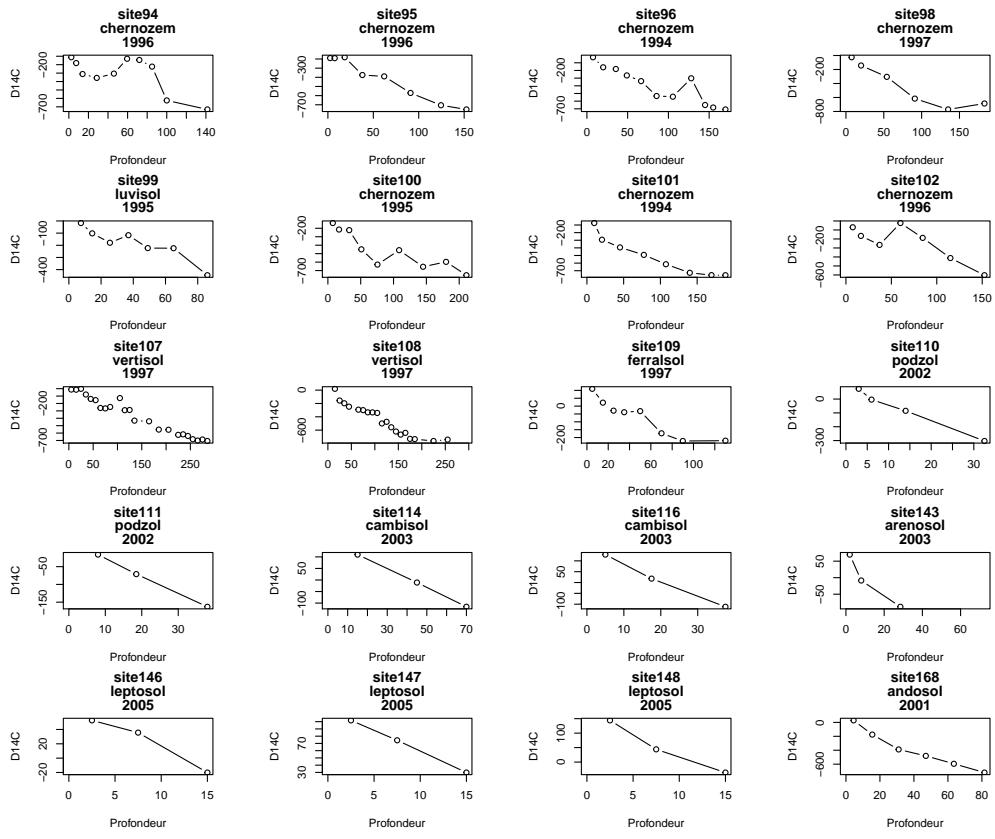
11 Perspectives

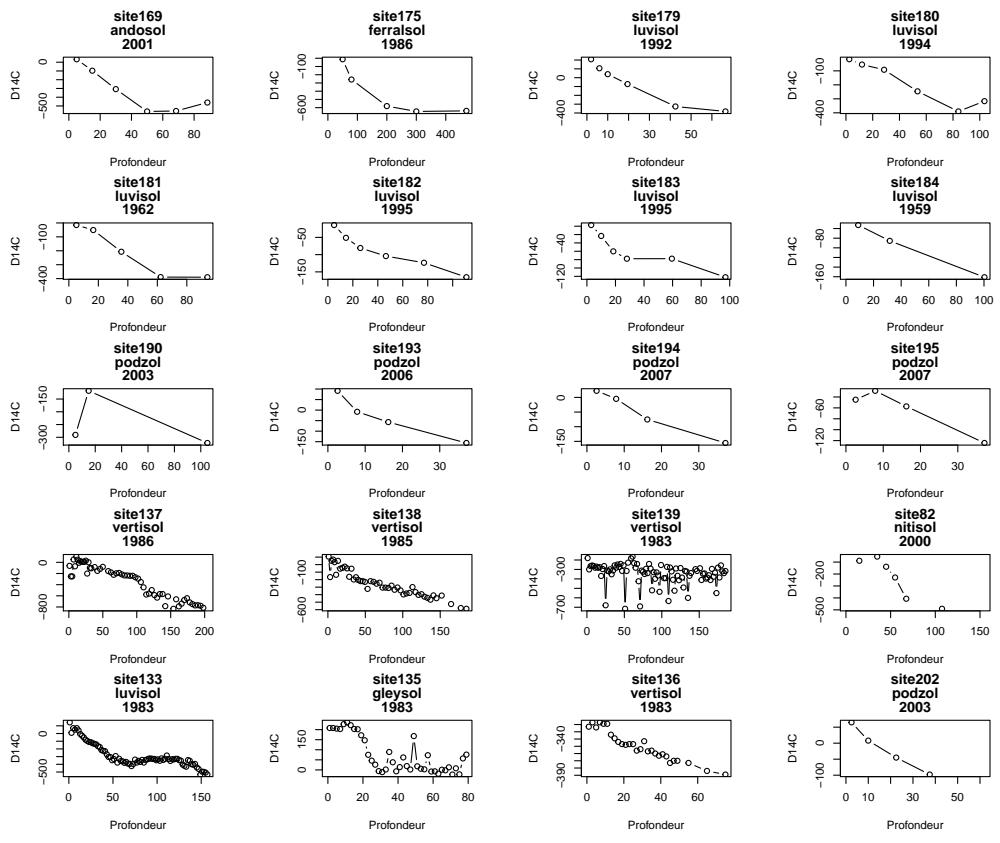
1. Tester l'algorithme VBEM (Variationnal Bayes Expectation Maximization), et le comparer avec les méthodes MCMC.
2. Considérer un modèle dont la variance varie avec l'épaisseur de la couche mesurée.
3. Appliquer l'algorithme EM de Jordane sur la nouvelle base de données afin de le comparer avec l'approche bayésienne.

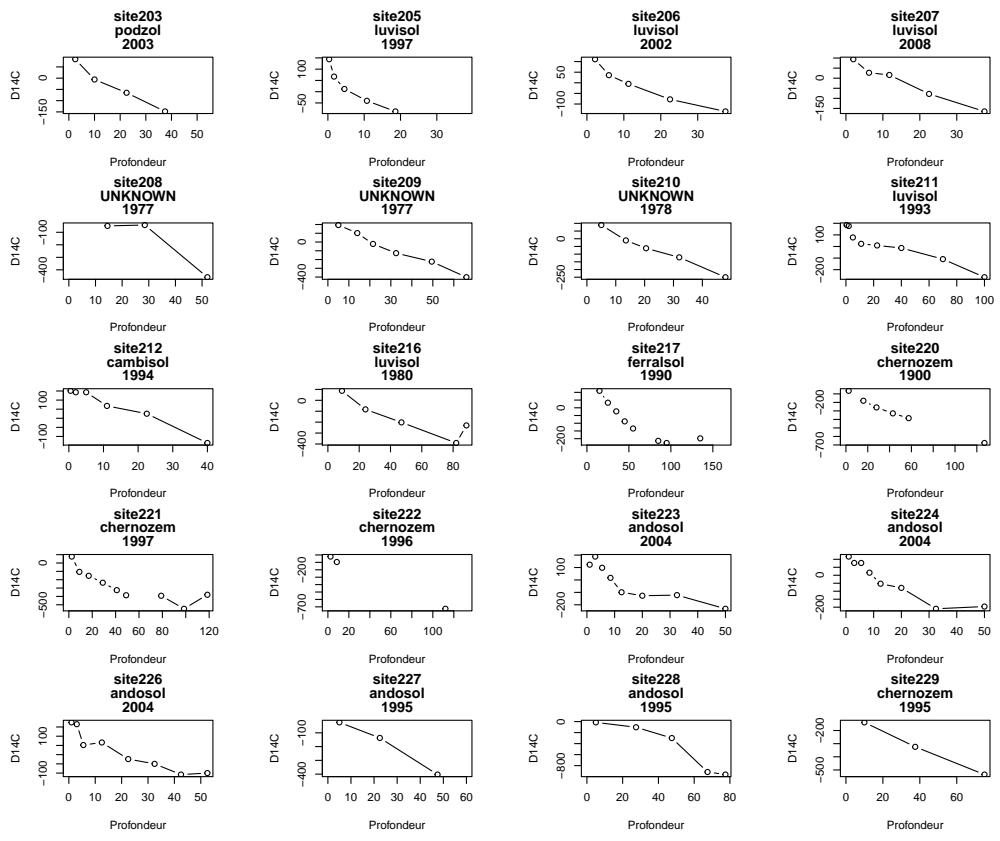
TO BE CONTINUED

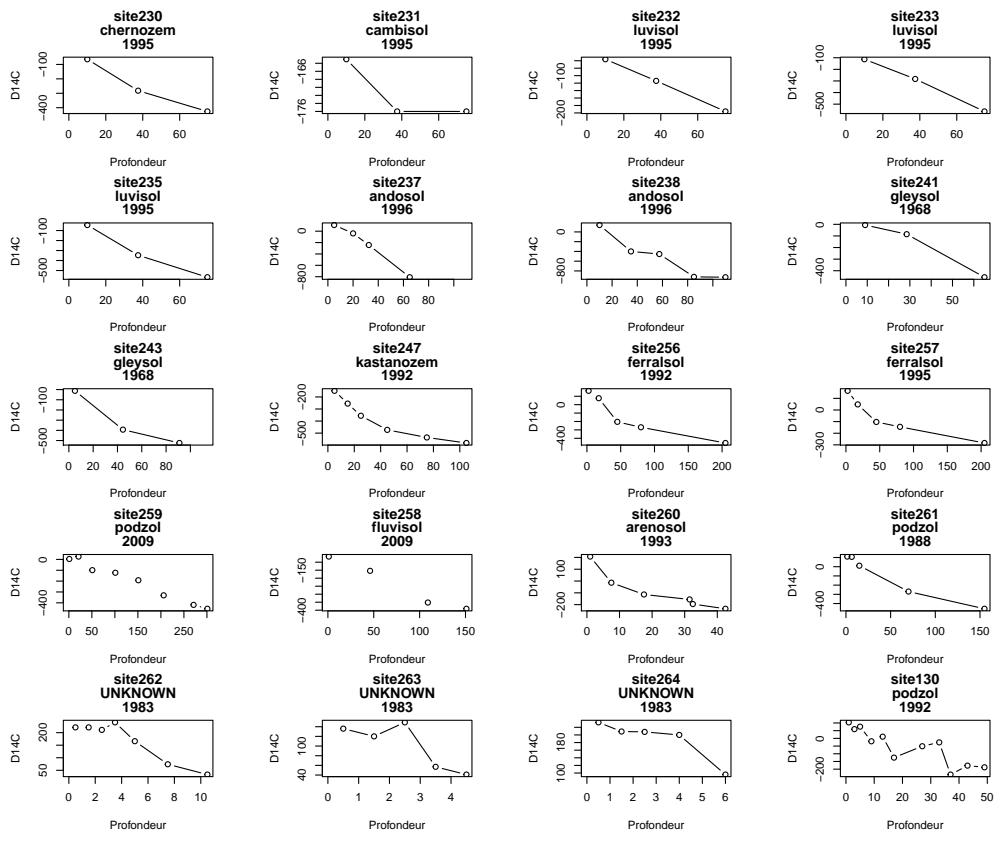
A Les profils sélectionnés pour l'analyse statistique

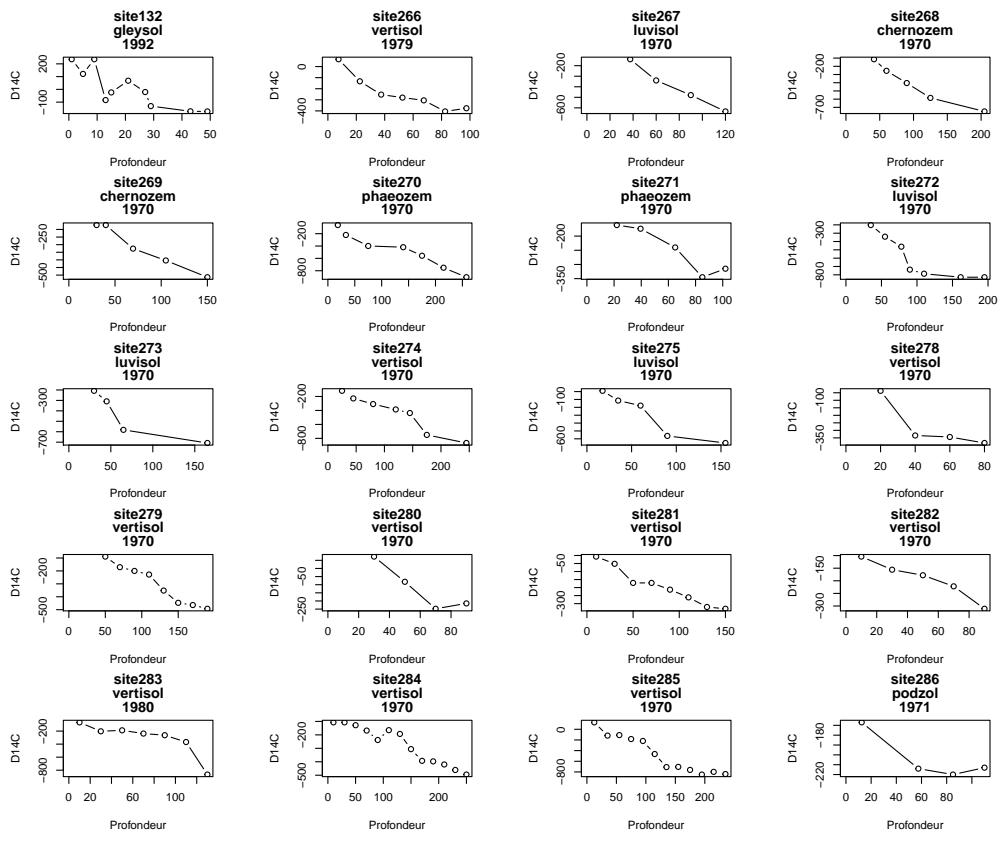


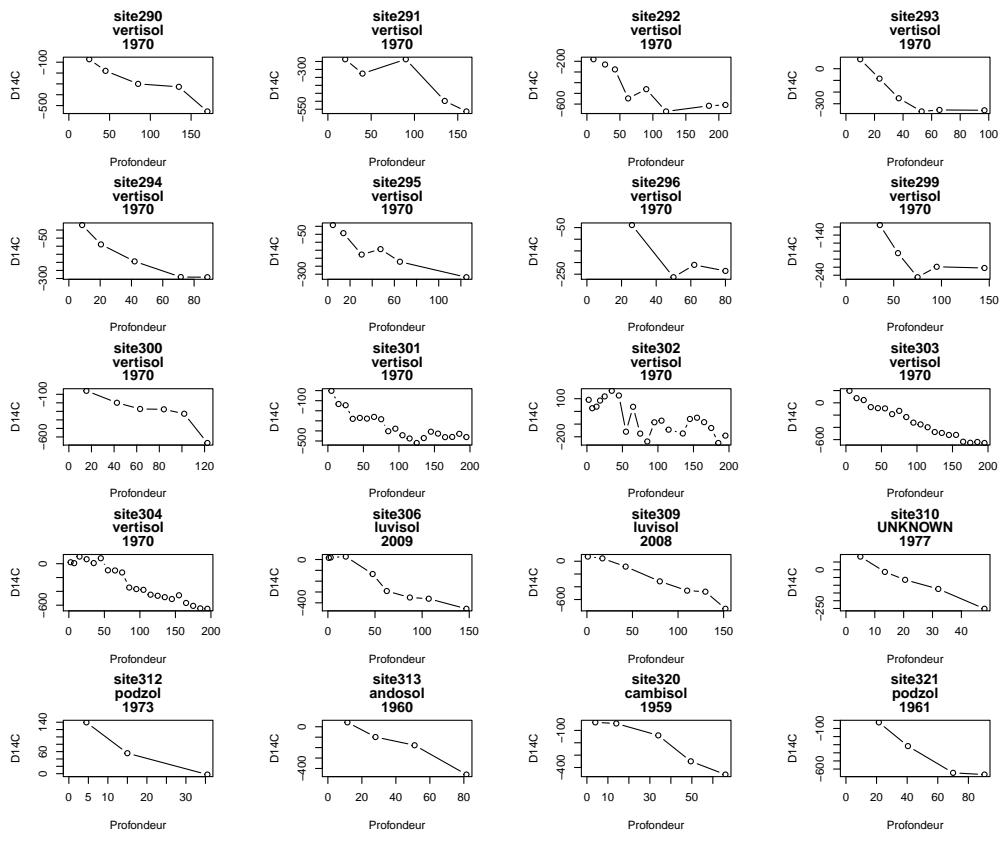


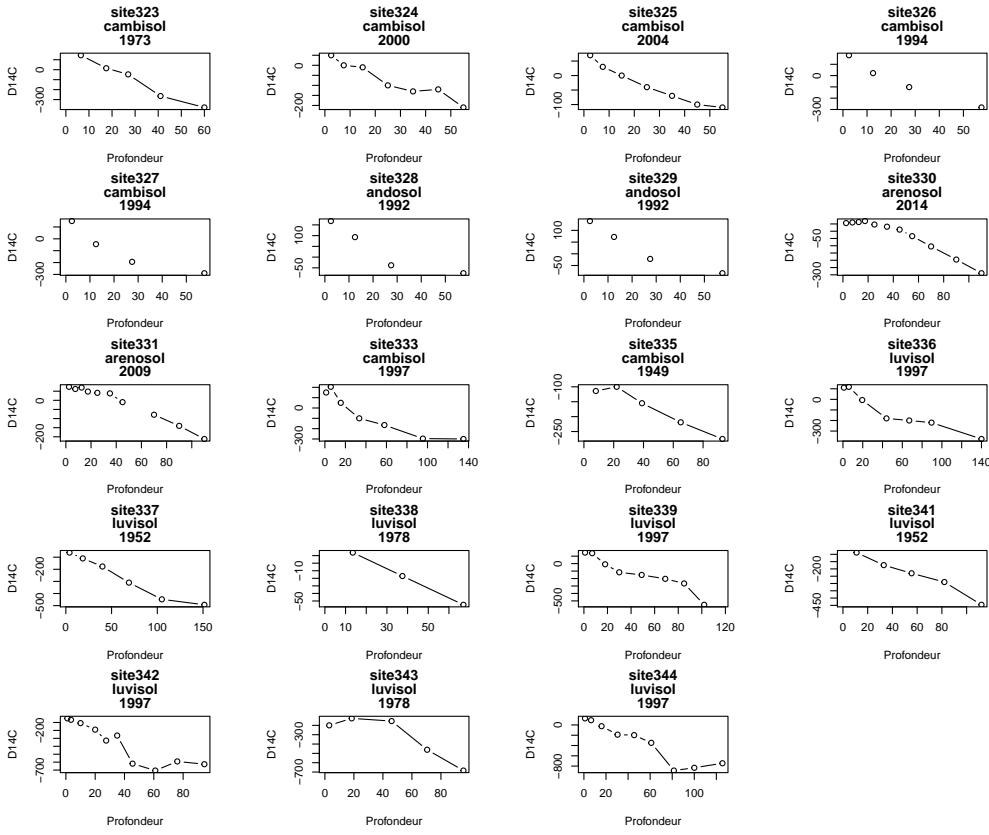












B Code R : Estimation via "optim"

```

#-----#
#      Ajustement et estimation
#-----#

# Modèle ajusté: premier travail c'est le modèle de Jordane

model <- function(phi,z){
  Y.ajus <- phi[1]+phi[2]*exp(-(z/phi[3])^phi[4])
  return(Y.ajus)
}

# fonction de coût à minimiser pour optim: ----->
#"somme des carrés des résidus"

func1 <- function(par,y,x) {
  sum((y-model(par,x))^2)
}

```

```

# Estimer phi1,phi2,phi3,phi4 en utilisant la fonction "optim"

PHI.optim = NULL
set.seed(123)

for (k in unique(Data$site_nb)) {
  v=which(Data$site_nb==k)
  x.top=Data$level_top[v]
  x.base=Data$level_base[v]
  x = 0.5*(x.top+x.base)
  y=Data$D14C[v]
  x=x[!is.na(y)] #on enlève les lignes sans D14C :
  y=y[!is.na(y)]

  palier=y[length(y)]
  surf=y[1]

  # xh:abscisse de la distance à mi-hauteur
  y.moy=(surf+palier)/2
  indi=which.min(abs(y-y.moy))
  xh=x[indi]

  a0=palier
  b0=(surf-palier)
  f0=2
  c0=xh*(log(2))^(-1/f0)

  est.o4 <- optim( c(a0, b0,c0,f0), func1,
                  method="SANN", hessian=TRUE, y=y, x=x)
  PHI.optim <- rbind(PHI.optim,est.o4$par)
}

PHI.optim <- cbind(unique(Data$site_nb),PHI.optim)
colnames(PHI.optim) <- c("site","phi1","phi2","phi3","phi4")
print(PHI.optim)

  site      phi1      phi2      phi3      phi4
[1,]    1 -544.536521  428.43684 100.900609  1.3765806
[2,]    2 -719.530612  595.15706 133.046799  2.7570968
[3,]    3 -299.436962  272.01582  79.647584  0.9148704
[4,]    7 -610.430429  687.04337  85.951637  2.2773376
[5,]   19 -411.830463  360.80867  55.084305  2.8506233
[6,]   20 -238.239800  377.99637  37.054990  1.0562613
[7,]   62 -732.692780  794.09976 116.614795  2.0781315
[8,]   63 -360.466432  443.28572  55.174220  1.5982244
[9,]   64 -444.526683  484.98629  65.565236  1.7870140
[10,]  66 -180.504709  298.14637  19.391959  1.1648590
[11,]  67 -334.892797  639.62690  25.634215  6.0489534
[12,]  68 -257.138489  504.69051  32.894704  1.3326151
[13,]  69 -354.269061  487.33155  93.013041  1.5854235
[14,]  70 -322.124002  603.71532  37.550639  1.0470394
[15,]  73 -359.056904  498.03707  27.647515  8.1794059
[16,]  74 -572.551270  651.89656 111.013312  1.8713780

```

[17,]	75	-501.512607	665.00589	34.828635	2.9638457
[18,]	91	-822.152303	791.99934	78.101413	1.7488792
[19,]	92	-774.928625	763.45892	81.471997	1.2666786
[20,]	93	-248.316056	247.50368	26.195344	0.6801471
[21,]	94	-734.385939	514.84631	98.736051	33.7518141
[22,]	95	-775.246248	588.03605	86.664501	1.8240065
[23,]	96	-719.449196	684.64119	93.872083	1.3050294
[24,]	98	-729.447947	663.48609	72.556118	2.5170783
[25,]	99	-519.482909	432.75696	76.924476	2.8763362
[26,]	100	-724.737210	636.92638	74.434696	1.0327817
[27,]	101	-778.501536	698.54494	72.710633	1.1842199
[28,]	102	-602.678866	468.28586	116.426662	7.8384092
[29,]	107	-738.154709	696.22868	167.934062	1.8186630
[30,]	108	-799.793971	728.77131	127.377748	1.8395030
[31,]	109	-247.167729	345.44586	57.512698	1.3297090
[32,]	110	-327.807939	392.67668	19.807496	1.8357484
[33,]	111	-163.056334	147.64095	21.120957	5.6727154
[34,]	114	-115.953501	225.92017	45.937486	12.8850235
[35,]	116	-112.036833	241.21117	18.966150	5.9862392
[36,]	143	-88.100567	165.24291	9.337738	2.0150302
[37,]	146	-20.040689	72.78718	9.671570	5.1782201
[38,]	147	29.847536	71.97132	7.996467	11.3901048
[39,]	148	-36.160596	180.99173	7.827835	4.7881823
[40,]	168	-720.466661	770.68697	37.478105	1.2912227
[41,]	169	-526.206426	548.14341	29.596080	2.4919053
[42,]	175	-621.973797	626.81577	88.032265	6.3678400
[43,]	179	-394.138609	601.47325	25.238863	1.3340202
[44,]	180	-352.698464	317.64149	50.717461	3.0178175
[45,]	181	-393.983778	375.91389	40.014289	2.9111019
[46,]	182	-170.230530	169.09097	46.824356	0.9503087
[47,]	183	-118.852954	145.88363	25.306256	0.7562822
[48,]	184	-160.977880	108.25158	37.034467	6.6282479
[49,]	190	-290.510313	70.46638	5.683422	-26.2396164
[50,]	193	-160.898839	280.07172	14.372424	1.1355564
[51,]	194	-158.193927	184.10369	18.096009	2.0317601
[52,]	195	-124.517269	87.67693	18.091600	11.9332326
[53,]	137	-779.647552	728.24721	119.987765	3.2726479
[54,]	138	-610.140791	684.55095	97.130135	0.9269338
[55,]	139	-372.598522	141.84717	31.572943	0.7004581
[56,]	82	-490.503713	419.79151	62.211638	6.0107830
[57,]	133	-518.132294	687.89579	51.325012	0.6980551
[58,]	135	21.088143	193.91301	22.698713	8.6427513
[59,]	136	-394.538361	78.50800	41.872840	1.2760835
[60,]	202	-113.761453	188.40981	20.522953	1.2862546
[61,]	203	-164.564939	259.31834	20.188411	1.2381896
[62,]	205	-87.970852	245.70063	5.087272	0.8482717
[63,]	206	-140.362759	271.69936	14.787091	1.1265646
[64,]	207	-188.150435	272.19394	23.227622	1.5799233
[65,]	208	-458.969268	413.51279	44.780617	27.3105325
[66,]	209	-412.200340	615.16523	38.643775	1.5990218
[67,]	210	-274.253607	368.66357	30.575676	1.5896657
[68,]	211	-284.550335	456.86837	56.611127	0.8427755
[69,]	212	-176.109423	319.53447	29.585380	1.9130982
[70,]	216	-300.375064	401.55120	34.708264	1.7705606
[71,]	217	-214.291135	352.94822	44.155866	2.0453751
[72,]	220	-705.683471	617.93854	69.102693	1.4257091

[73,]	221	-434.587494	478.42748	28.562930	1.2260120
[74,]	222	-726.008980	704.09050	13.460927	4.8084381
[75,]	223	-166.771296	330.08075	10.499232	2.5770699
[76,]	224	-209.512346	323.57086	17.681648	1.4666773
[77,]	226	-114.251327	294.85890	18.888420	1.0214440
[78,]	227	-403.086019	379.08684	25.015033	9.8035679
[79,]	228	-967.939594	917.01529	57.143293	6.1226159
[80,]	229	-534.951489	397.09155	40.835408	5.3727630
[81,]	230	-425.988972	363.50379	38.206941	4.1054644
[82,]	231	-176.003570	13.02123	20.069965	9.8357204
[83,]	232	-196.067001	140.04725	40.774745	7.4839083
[84,]	233	-562.995522	450.33247	43.724158	4.8482029
[85,]	235	-567.042597	525.92807	38.954797	3.6130233
[86,]	237	-835.036600	927.55717	42.689030	2.7493696
[87,]	238	-945.881784	1099.09331	52.066678	1.6341150
[88,]	241	-455.952520	450.05934	37.769110	5.8384432
[89,]	243	-523.945935	511.92504	41.414530	4.3166376
[90,]	247	-572.265914	458.05619	32.254797	1.2844441
[91,]	256	-450.350363	640.14597	56.647652	1.1460007
[92,]	257	-286.395068	473.67392	54.923393	0.8947137
[93,]	259	-487.037370	480.85831	195.714463	1.9594245
[94,]	258	-396.723630	331.80429	78.833900	2.1568507
[95,]	260	-221.123208	464.17643	10.757619	0.8888616
[96,]	261	-471.205479	587.10413	64.465178	1.2756265
[97,]	262	34.452038	190.21342	6.674783	4.3486991
[98,]	263	40.824600	94.61316	3.407999	21.1188134
[99,]	264	83.597499	133.57997	5.295294	6.8231421
[100,]	130	-201.404133	304.43066	22.984155	1.0864600
[101,]	132	-186.398293	416.30046	21.225866	1.1996250
[102,]	266	-371.448127	478.63803	31.439902	1.3400973
[103,]	267	-634.401911	560.03390	76.575374	2.4844709
[104,]	268	-744.792399	684.76497	104.627094	2.2580445
[105,]	269	-521.070433	394.73590	89.334943	2.1042591
[106,]	270	-963.512849	829.58465	186.362372	1.8240930
[107,]	271	-329.916792	162.83165	68.644599	9.4546475
[108,]	272	-816.490052	554.23246	84.286042	9.2063076
[109,]	273	-707.809846	521.50685	60.133240	4.5359625
[110,]	274	-882.039149	682.54060	163.370627	4.1089448
[111,]	275	-655.770512	617.35997	78.828916	4.5718561
[112,]	278	-363.992395	342.82464	34.200812	5.6832408
[113,]	279	-496.698504	387.78246	131.001024	4.0636681
[114,]	280	-231.582872	304.82714	51.305475	13.1991206
[115,]	281	-353.783543	353.86412	84.844527	1.5159147
[116,]	282	-367.584490	249.21666	81.802023	2.6676312
[117,]	283	-958.906722	774.39636	122.148459	11.4735572
[118,]	284	-514.630832	393.99980	178.452282	2.9280365
[119,]	285	-868.666621	919.90697	123.080964	2.4191464
[120,]	286	-216.532424	52.51491	37.449312	2.6065540
[121,]	290	-577.910412	503.57304	125.787999	1.7341338
[122,]	291	-563.969393	297.86961	133.018010	27.6902527
[123,]	292	-618.413622	460.37667	62.507344	2.0352787
[124,]	293	-360.202130	467.57788	30.900241	2.4415314
[125,]	294	-297.400043	361.41606	33.652268	1.4655349
[126,]	295	-328.644090	353.60027	52.591571	1.0482801
[127,]	296	-236.344589	197.40149	41.569496	17.4596487
[128,]	299	-228.691351	114.81133	54.675771	10.6489197

```

[129,] 300 -788.714818 606.78318 116.230080 7.7652163
[130,] 301 -472.328543 467.74245 55.437615 1.1195755
[131,] 302 -220.091254 306.38326 115.092644 1.2670169
[132,] 303 -701.259525 835.49488 109.493423 1.6234439
[133,] 304 -621.957544 691.36608 104.722951 2.3116008
[134,] 306 -429.986165 460.19674 65.074569 2.1250449
[135,] 309 -774.871423 834.21069 108.997328 1.9155970
[136,] 310 -275.190028 366.14071 30.397687 1.6031472
[137,] 312 -2.399636 142.00608 15.185638 9.1772508
[138,] 313 -518.924894 529.62651 64.200901 2.6771181
[139,] 320 -462.374007 424.35098 46.264215 4.1743638
[140,] 321 -657.328704 633.12478 42.338777 10.7286159
[141,] 323 -397.320492 536.86201 36.212089 2.2866163
[142,] 324 -221.236612 270.04018 33.936205 1.3382085
[143,] 325 -127.412388 202.56620 27.665609 1.2243969
[144,] 326 -299.186463 485.94462 27.251626 1.3116278
[145,] 327 -286.615167 468.45342 17.912228 1.2459569
[146,] 328 -74.942480 246.71550 20.051505 2.0199573
[147,] 329 -87.517804 235.61644 23.348335 1.5070036
[148,] 330 -319.907752 380.93132 84.504720 3.0213549
[149,] 331 -245.082215 310.61184 83.670571 2.3482315
[150,] 333 -300.531371 485.76839 40.949535 1.3088061
[151,] 335 -286.868077 182.13351 64.316056 2.6691436
[152,] 336 -375.855806 508.31957 58.138520 1.0462432
[153,] 337 -503.842613 433.44169 75.414827 1.9653357
[154,] 338 -55.011884 69.02309 40.571606 6.5729830
[155,] 339 -573.982563 710.65201 75.725051 1.2740079
[156,] 341 -493.835534 389.36191 88.850519 2.1550428
[157,] 342 -646.843770 565.51106 37.678108 3.1048058
[158,] 343 -683.721243 524.01491 71.001624 21.7563212
[159,] 344 -825.571621 848.39173 63.973483 3.8217449

```

les profils ajustés par "optim" sont représentés en rouge :

```

#Tracer les profils ajustés

par(mfrow = c(5,4),mar = c(2,2,2,2))

for (i in 1:length(unique(Data$site_nb))){

  S = Data[Data$site_nb == unique(Data$site_nb)[i]
  ,c("site_nb","level_base","level_top","D14C","sampling_year")]

  S = na.exclude(S)
  x.top <- S$level_top
  x.base <- S$level_base
  x <- 0.5*(x.top+x.base)
  y <- S$D14C

  phi.optim <- PHI.optim[i,-1]

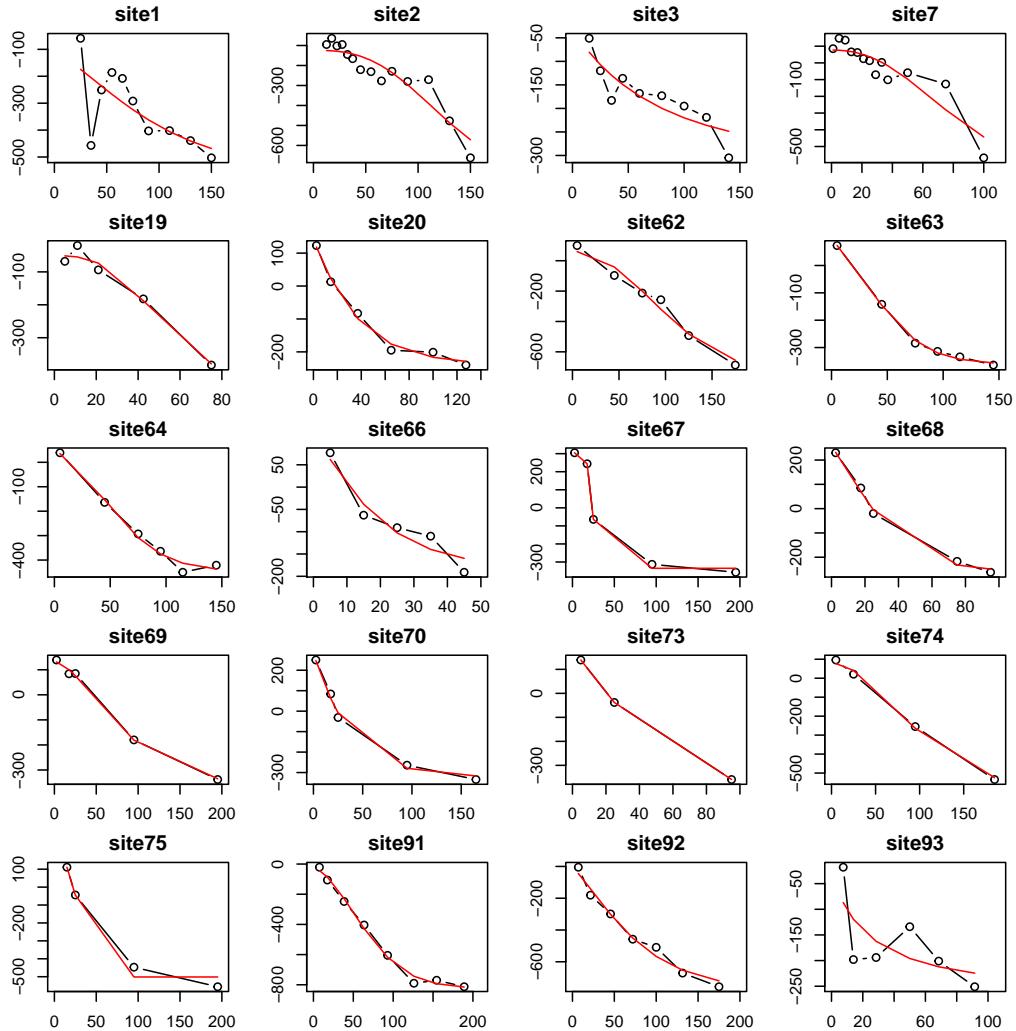
  plot(x,y,type = "b",xlab ="Profondeur",ylab ="D14C",
    main = c(paste("site",unique(Data$site_nb)[i],sep = ""),
    unique(S$samp1_year)),xlim =c(0,max(S$level_base,na.rm =T)))
}

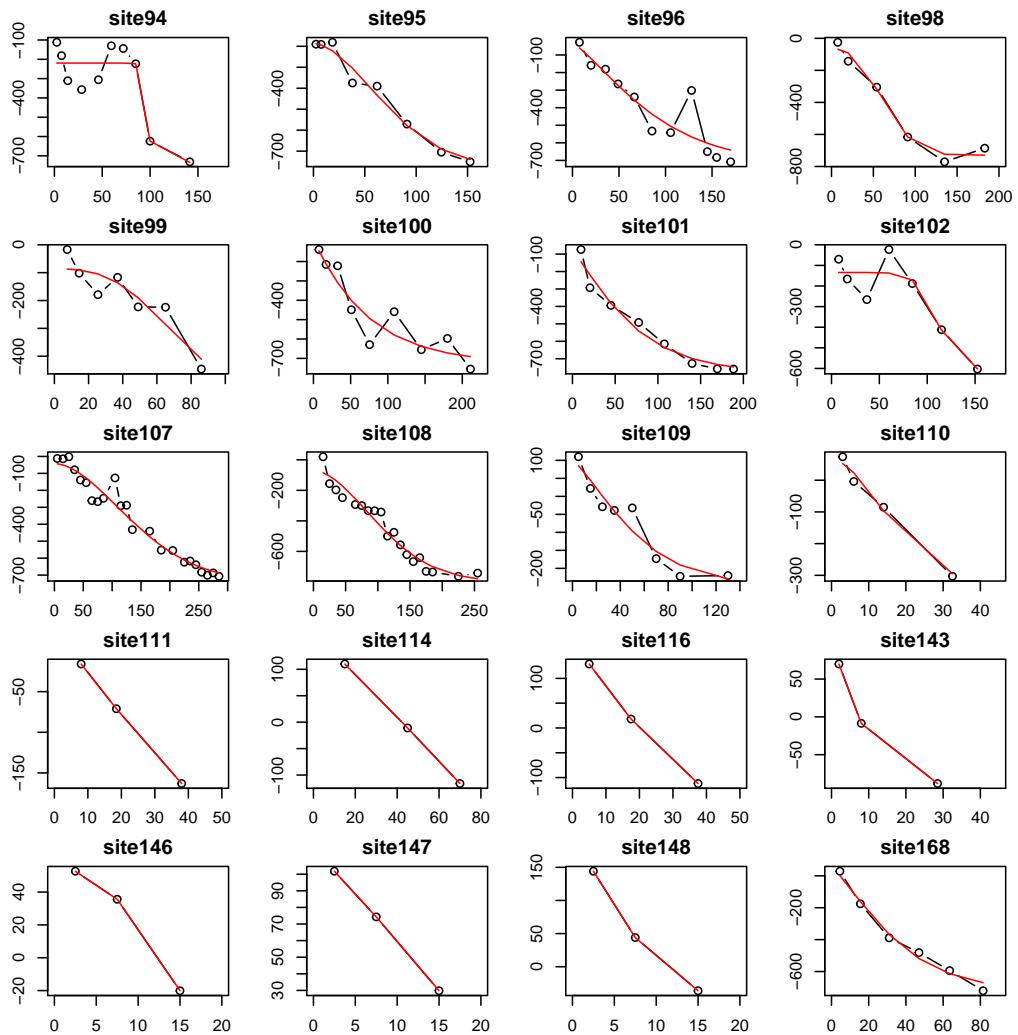
```

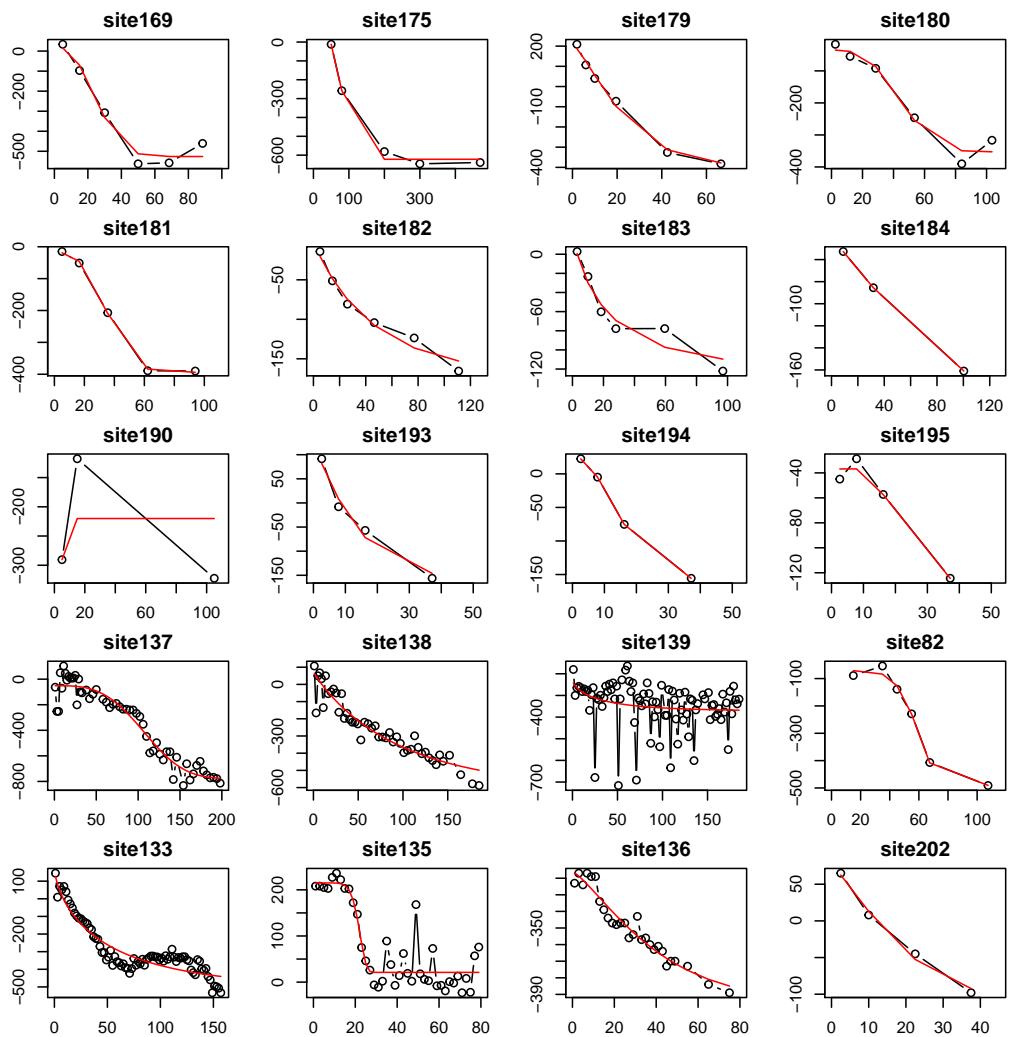
```

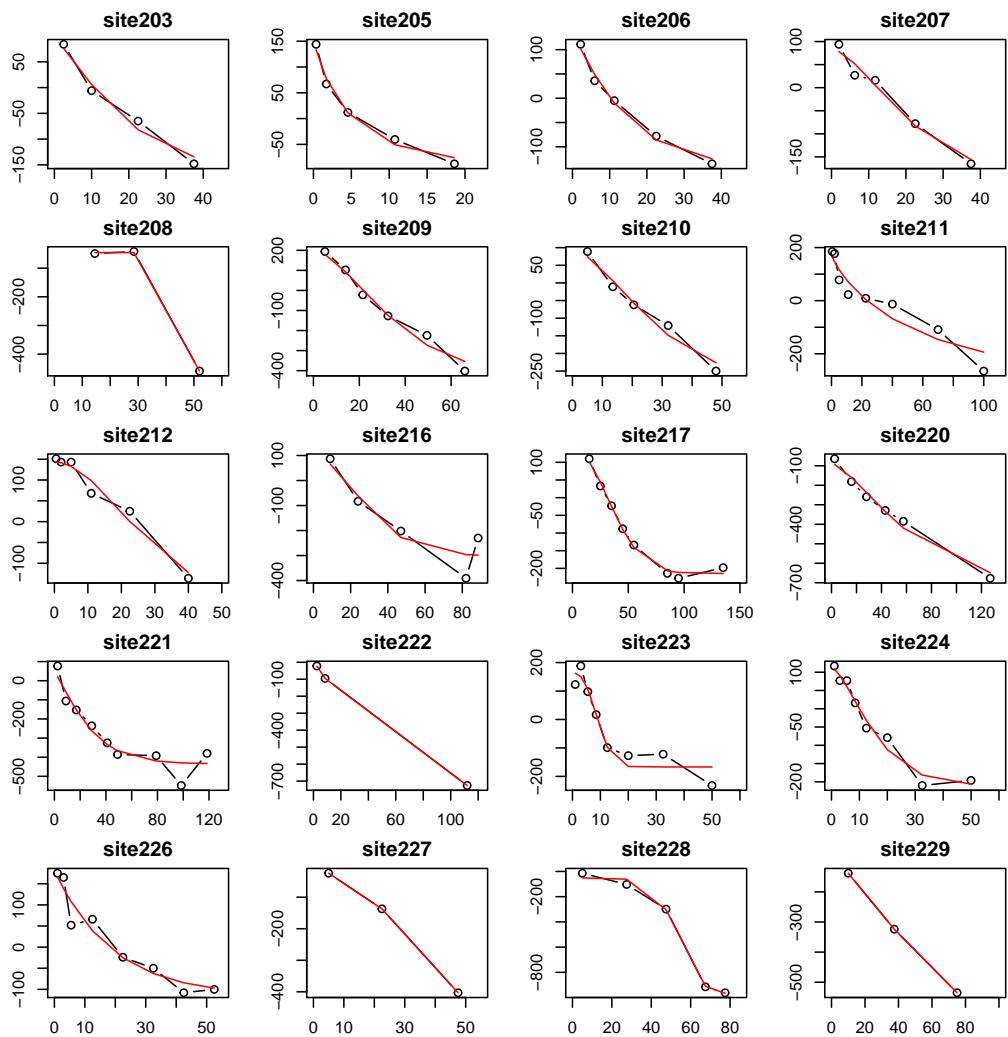
points(x,model(phi = phi.optim,x),col = "red",type ="l")
}

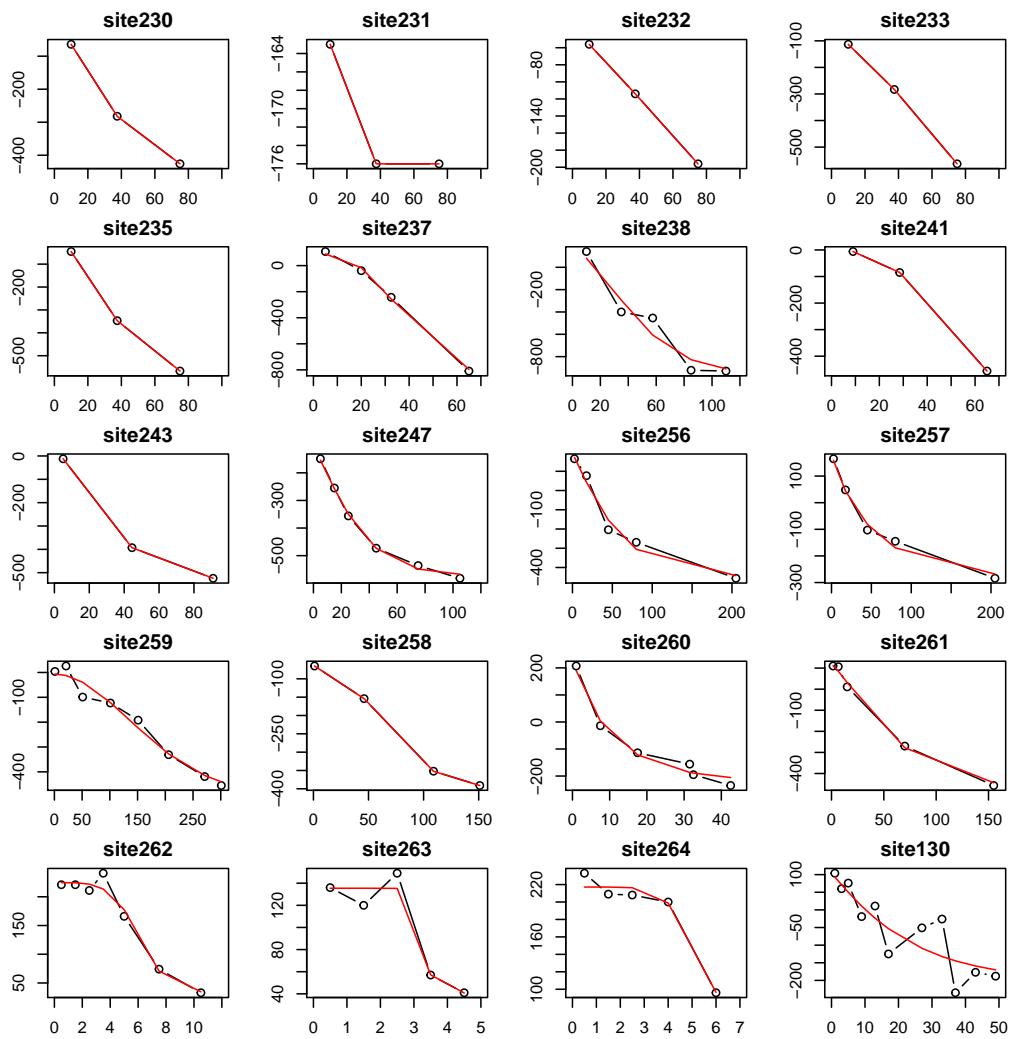
```

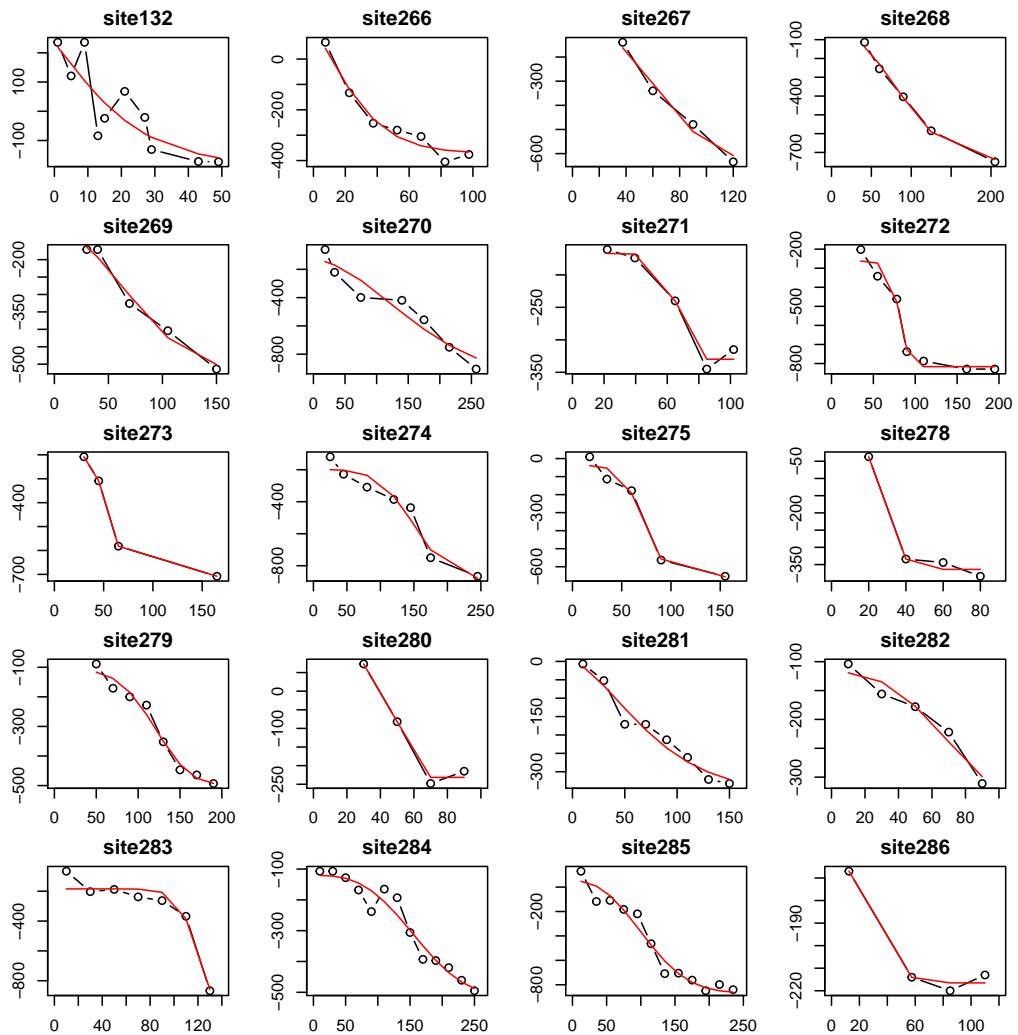


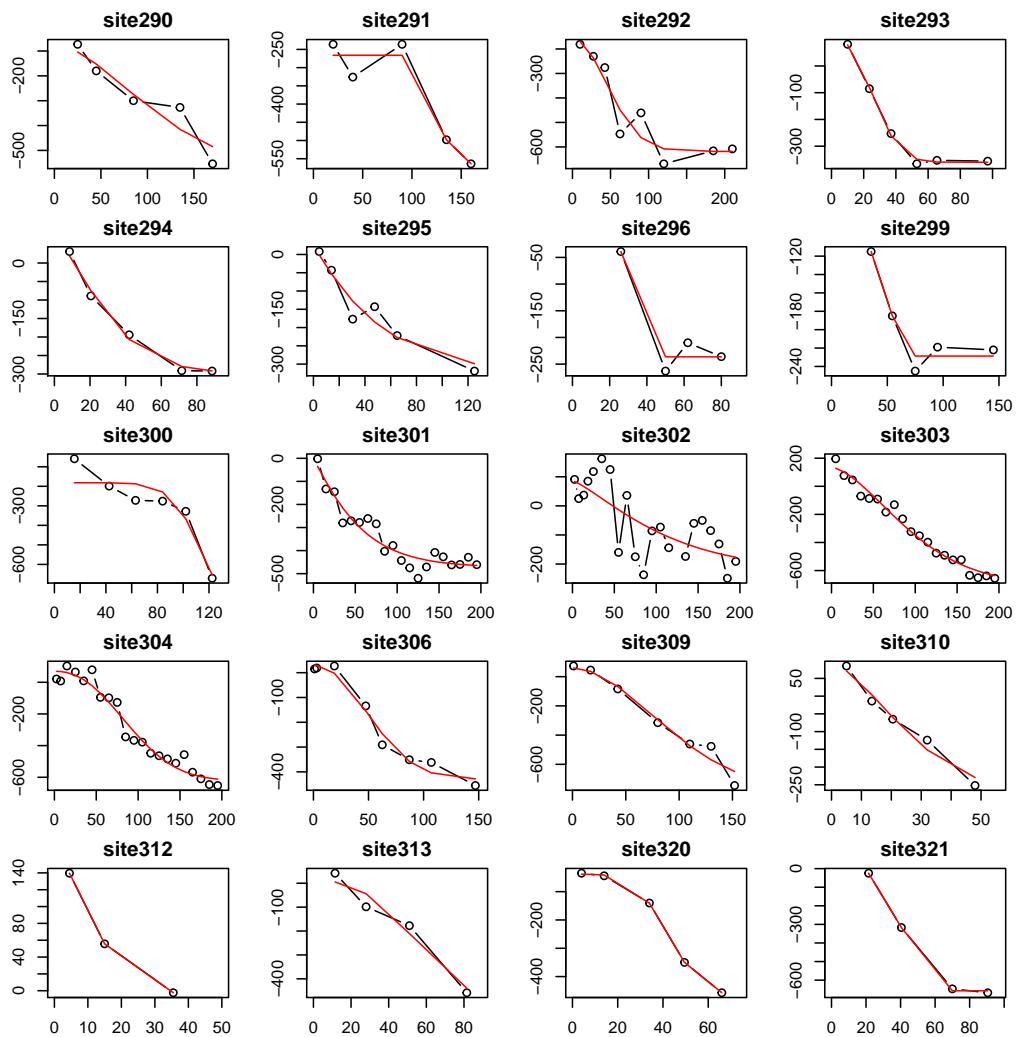


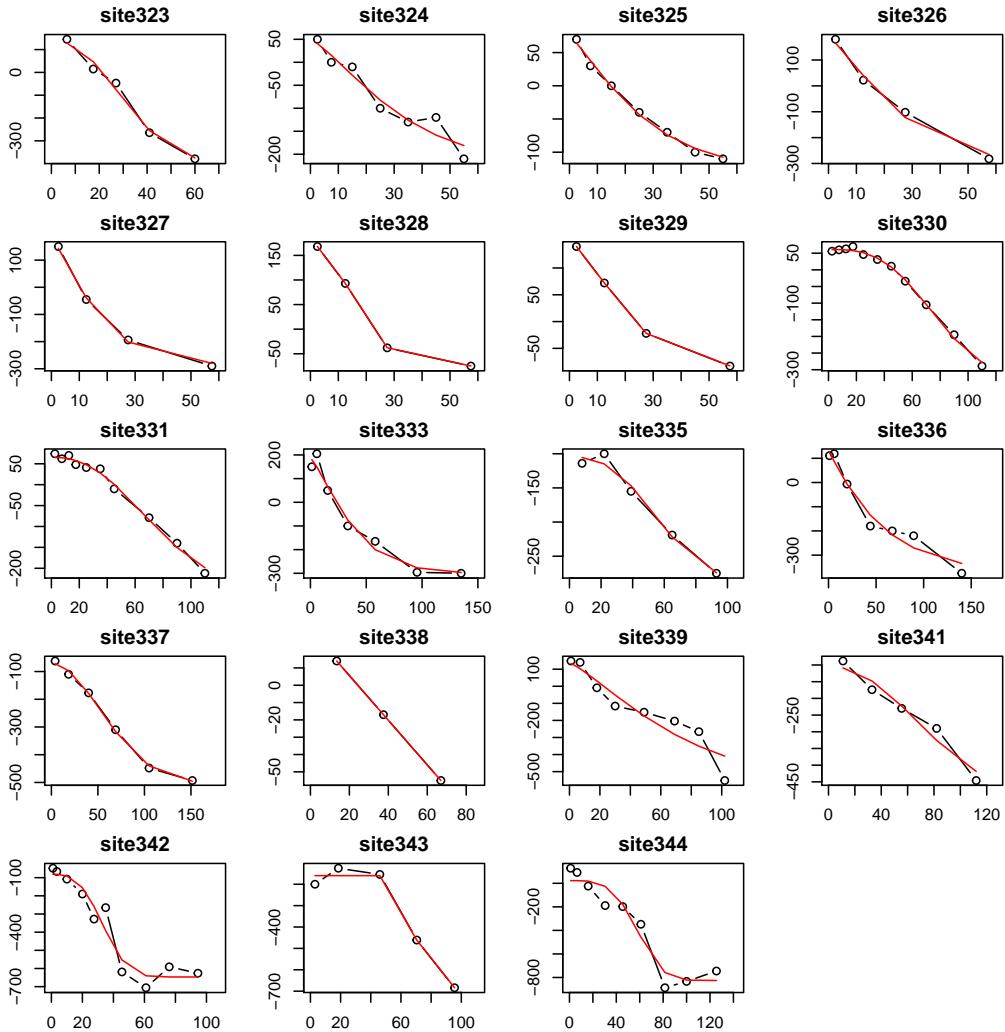












C Code R : Nettoyage des données

```

#-----
#   Préparation de la base de données
#-----

# exclu le site 190 ayant la valeur de phi4 négatif
# exclu les sites 146, 147, 148, 238
Data.selected = Data[Data$site_nb %!in% c(190, 146, 147, 148, 238), ]

```

```

# ici le type de sol inconnu est noté par UNKNOWN
# trouver les indices dont le type de sol est UNKNOWN afin de les supprimer

site_UNKNOWN_WRB= Data.selected[Data.selected$WRB == "UNKNOWN"
                                , "site_nb"]
Data.selected=Data.selected[Data.selected$site_nb%in%
                                unique(site_UNKNOWN_WRB),]

#trouver les indices dont le type de landuse est NA afin de les #supprimer

site_Na_landuse = Data.selected[Data.selected$landuse == "NA",
                                "site_nb"]
Data.selected=Data.selected[Data.selected$site_nb %!in%
                                unique(site_Na_landuse),]

# Supprimer les valeurs manquantes pour les variables numériques
Data.selected = na.exclude(Data.selected)

## voir le nbr de profils par WRB
hh=unstack(Data.selected,form=
            Data.selected$site_nb~Data.selected$WRB)
nb_profil.par.WRB=sapply(hh
                           ,function(x){length(unique(unlist(x)))})

## voir le nbr de profils par type de landuse
dd = unstack(Data.selected,form=
            Data.selected$site_nb~Data.selected$landuse)
nb_profil.par.landuse = sapply(dd
                               ,function(x){length(unique(unlist(x)))})

-----
#Préparation des variables explicatives quantitatives
-----

site_a_choisir = unique(Data.selected$site_nb)
ind <- which(PHI.optim[,1] %in% site_a_choisir)
PHI.optim.selected <- PHI.optim[ind,]

{
  Atm=c()
  Sol=c()
  Land=c()
  Stock_surf=c()
  Temp=c()
  Lat=c()
  Ann=c()
  Samp=c()
  Stock_prof=c()
  Pann=c()
  Arid=c()
  Prof_max=c()
  T_janv=c()
  Longi=c()
  Alti=c()
  T_ju=c()
}

```

```

P_janv=c()
P_ju=c()
Dif_T=c()

k1=1
for (k in PHI.optim.selected[,1]) {
t=which(Data.selected$site_nb==k)
x2=0.5*(Data.selected$level_base[t]+Data.selected$level_top[t])

#indi=which(param_ok[,1]==k)

sol=Data.selected$WRB[t[1]]
Sol=rbind(Sol,sol)

land=Data.selected$landuse[t[1]]
Land=rbind(Land,land)

Samp=rbind(Samp,Data.selected$sampling_year[t[1]])
Atm=rbind(Atm,Data.selected$D14Catm[t[1]])

stock= Data.selected$level_stock[t]
stock=stock/(Data.selected$level_base[t]-Data.selected$level_top[t])
Stock_surf=rbind(Stock_surf,stock[1])
Stock_prof=rbind(Stock_prof,stock[length(stock)])

Temp=rbind(Temp,Data.selected$Tann[t[1]])
Lat=rbind(Lat,Data.selected$latitude[t[1]])
Longi=rbind(Longi,Data.selected$longitude[t[1]])
Alti=rbind(Alti,Data.selected$altitude[t[1]])
Pann=rbind(Pann,Data.selected$Pann[t[1]])
Arid=rbind(Arid,Data.selected$aridity[t[1]])
Prof_max=rbind(Prof_max,x2[length(x2)])
T_janv=rbind(T_janv,Data.selected$Tjan[t[1]])
T_ju=rbind(T_ju,Data.selected$TJul[t[1]])
P_janv=rbind(P_janv,Data.selected$Pjan[t[1]])
P_ju=rbind(P_ju,Data.selected$PJJul[t[1]])
Dif_T=rbind(Dif_T
,abs(Data.selected$Tjan[t[1]]-Data.selected$TJul[t[1]]))
k1=k1+1
}

#-----
#   Préparation des variables explicatives qantitatives
#-----

Fp.total <- NULL
Fp.total=rbind(Fp.total,cbind(Atm,Temp,Pann,Lat,Arid,Stock_surf,Stock_prof,Dif_T))
Fp.total=matrix(as.vector(Fp.total),ncol=length(Fp.total[1,]))
colnames(Fp.total) <- c("Atm","Temp","Pann","Lat","Arid",
"Stock_surf","Stock_prof","Dif_T")
which(is.na(Fp.total)) # pas de valeurs manquantes

integer(0)

# regrouper fluvisol,kastanozem, phaeozem en un seul groupe sous le nom #"flu-

```

```

visol"
indice_WRB = which(Sol[,1] %in% c(6,8,12))
Sol = replace(as.vector(Sol),list = indice_WRB,values = 6)
type_sol<-lapply(Sol,function(x)levels(Data$WRB)[x])
type_sol <- as.factor(unlist(type_sol))
# regrouper natural-desert et natural en une seule catégorie sous le nom "natural"
indice_land <- which(Land[,1] %in% c(8,7))
Land <- replace(as.vector(Land),list=indice_land,values = 7)
type_landuse<-lapply(Land,function(x)levels(Data$landuse)[x])
type_landuse<-as.factor(unlist(type_landuse))

# Regrouper les variables quantitatives et qualitatives dans un seule data frame
df <- data.frame(type_landuse,type_sol,Fp.total)
p = ncol(df)
S = nrow(df)
# Appliquer model.matrix pour bien identifier entre les 2 types de variables
model.mat <-model.matrix(~ type_landuse+type_sol+scale(Atm)
+scale(Temp)+scale(Pann)+scale(Lat)+scale(Arid)+
scale(Stock_surf)+scale(Stock_prof)+scale(Dif_T),
data = df )

```

Les 55 sites éliminés sont représentés ci-dessous :

```

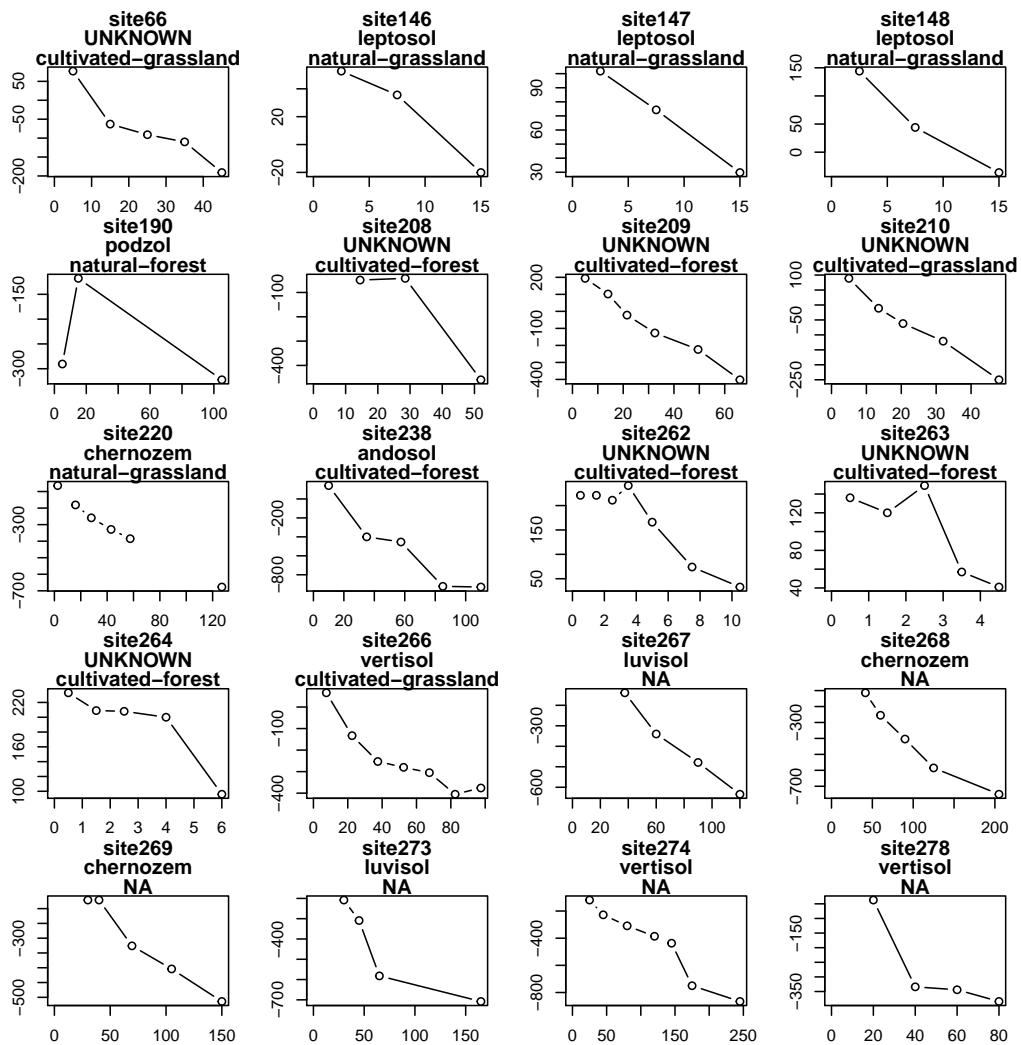
site.total = unique(Data$site_nb)
site.learning = unique(Data.selected$site_nb)
site.validation = site.total[site.total %!in% site.learning]

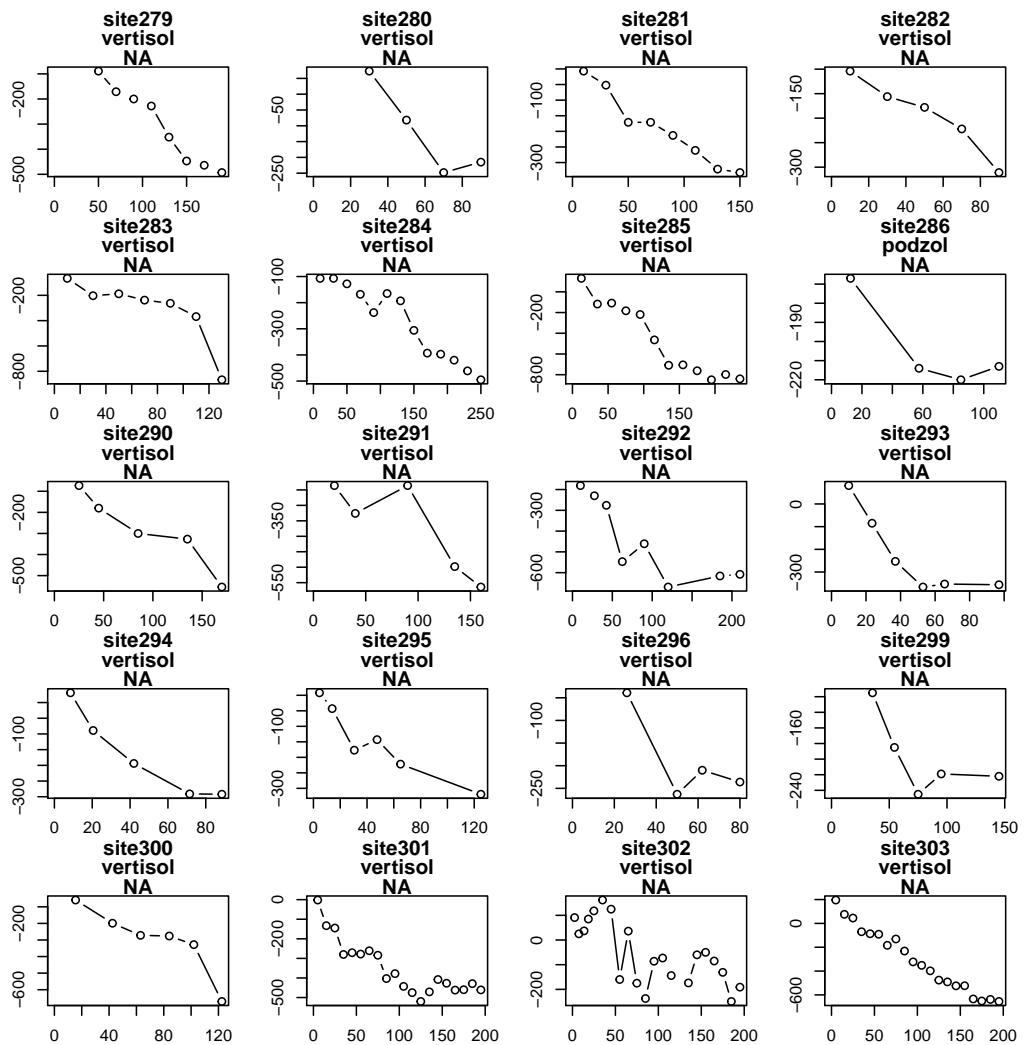
par(mfrow = c(5,4),mar = c(2,2,3,2))
for (i in site.validation){
  SS = Data[Data$site_nb == i ,]
  SS$level_mean <- 0.5* (SS$level_top+SS$level_base)

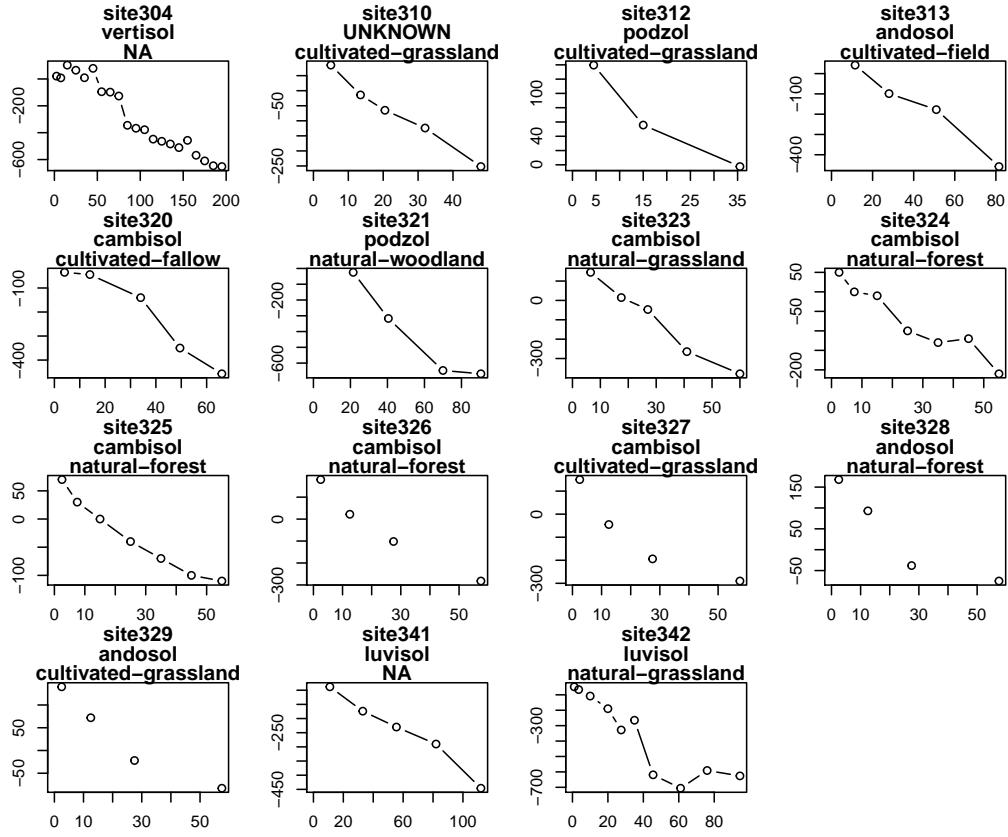
  type.sol = levels(Data$WRB)[unique(SS$WRB)]
  type.ecosys = levels(Data$landuse)[unique(SS$landuse)]
  year = unique(SS$sampling_year)

  plot(SS$level_mean,SS$D14C,type = "b",xlab ="Profondeur",
  ylab ="D14C",main = c(paste("site",i,sep=""),type.sol,
  type.ecosys),xlim = c(0,max(SS$level_mean,na.rm = T)))
}

```







D Code R :Manova et données simulées

```

#-----
#      Manova
#-----

# La réponse: phi1(s),phi2(s),phi3(s),phi4(s)
PHI <- PHI.optim.selected[, -1]

#Pour des raisons de positivité
PHI[, 3] <- log(PHI[, 3])

```

```

PHI[,4] <- log(PHI[,4])

Res1 <- manova(PHI~ model.mat-1)
summary(Res1) # les va.explicitives sont significatives

##          Df Pillai approx F num Df den Df    Pr(>F)
## model.mat 26   2.362   4.3261     104     312 < 2.2e-16 ***
## Residuals 78
## ---
## Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1

theta.hat <-Res1$coefficients
#Res1$residuals
Omega.hat <- var(Res1$resid)
#cor(Res1$resid)

#-----
#   Données simulées
#-----

# la matrice des variables latentes
set.seed(123)
phi.sim <- matrix(NA,nrow = S,ncol =ncol(PHI.optim.selected)-1)

for (i in 1:S) {
  mu = model.mat[i,] %*% theta.hat
  phi.sim[i,] <- mvrnorm(1,mu,Omega.hat)
}

# les données de D14C simulées
set.seed(123)
y.sim = list()
z = list()
for (i in 1:S){
  SS = Data.selected[Data.selected$site_nb == unique(Data.selected$site_nb)][i,]
  SS = na.exclude(SS)
  x <- 0.5*(SS$level_top +SS$level_base)
  y <- SS$D14C

  z[[i]] = zz = x
  n = length(x)
  phi <- phi.sim[i,]
  #phi[2] <- exp(phi[2])
  phi[3] <- exp(phi[3])
  phi[4] <- exp(phi[4])
  y.sim[[i]] = rnorm(n,model(phi,zz),sqrt(100))
}

# Visualisation des données simulés

par(mfrow =c(5,4),mar = c(2,2,2,2))
for (i in 1:104) {
  SS = Data.selected[Data.selected$site_nb == unique(Data.selected$site_nb)][i,]
  SS = na.exclude(SS)

```

```

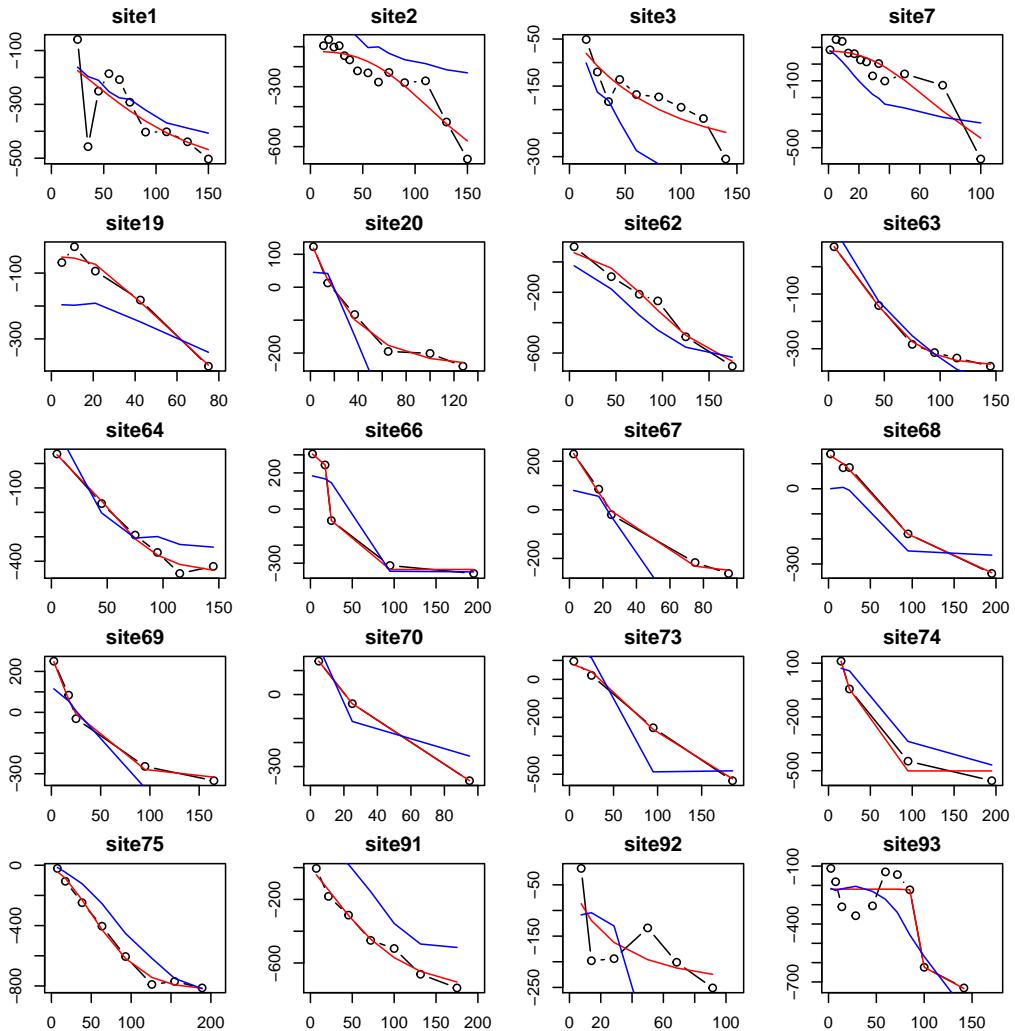
x <- 0.5*(SS$level_top+SS$level_base)
y <- SS$D14C

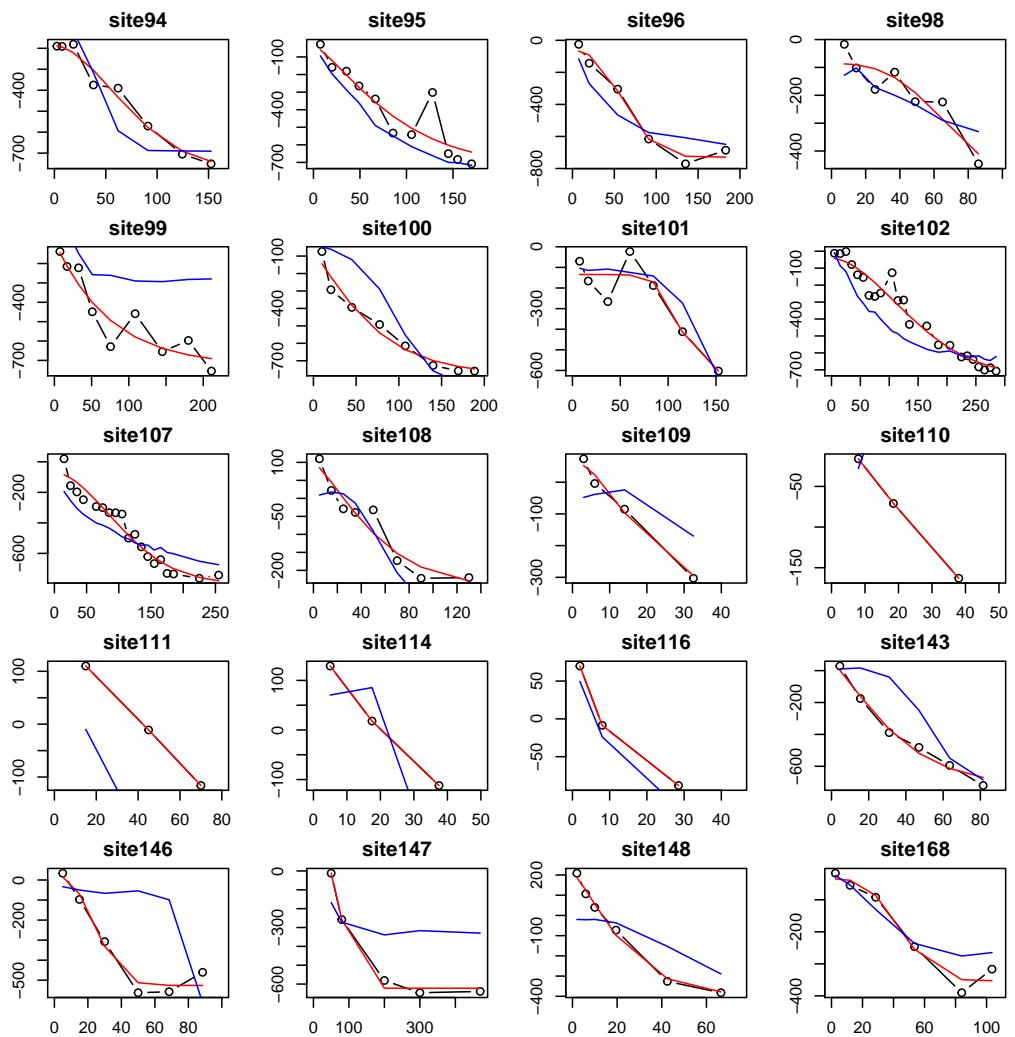
plot(x,y,type ="b",xlab ="Profondeur",ylab ="D14C",
main = c(paste("site" ,unique(Data$site_nb)[i] ,sep ="")),
xlim = c(0,max(SS$level_base,na.rm = T)))

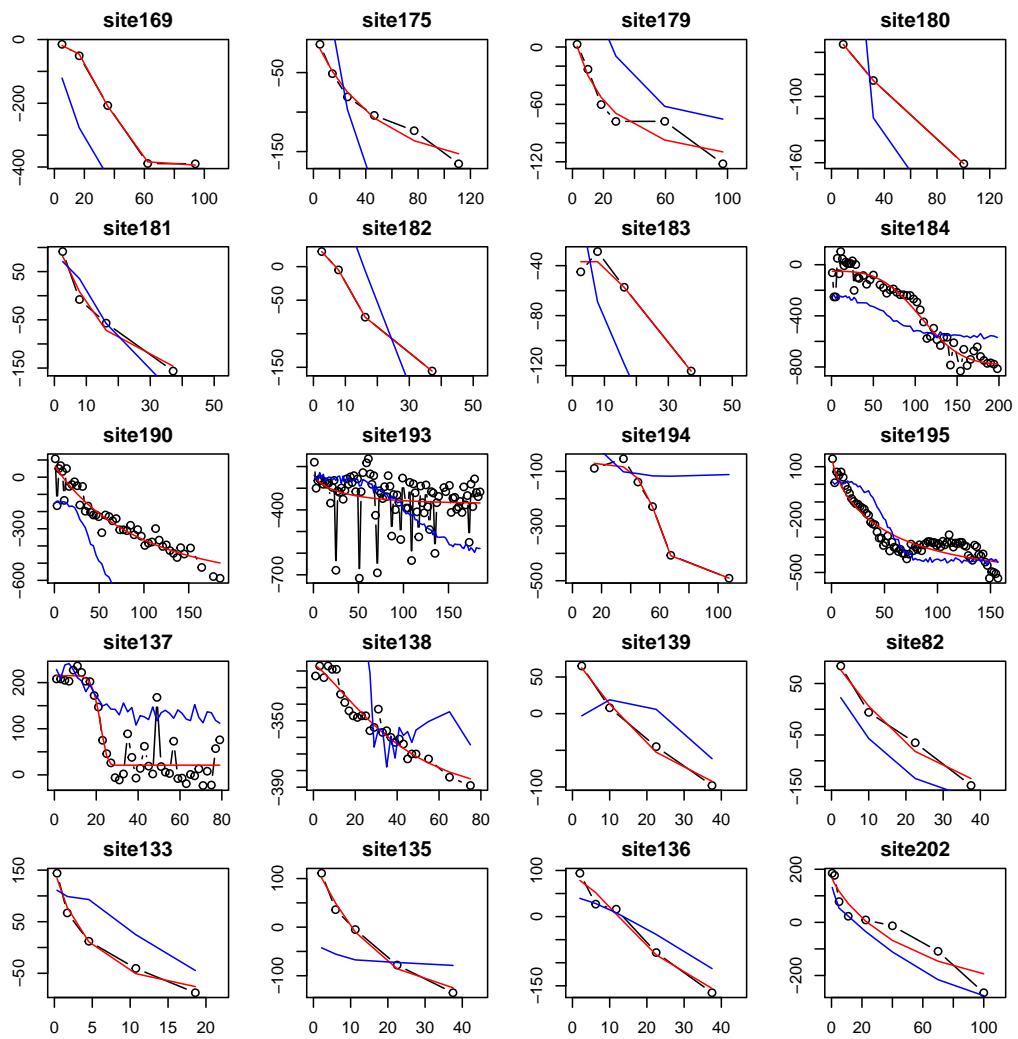
phi.optim <- PHI.optim.selected[i,-1]
y.simul <- y.sim [[i]]

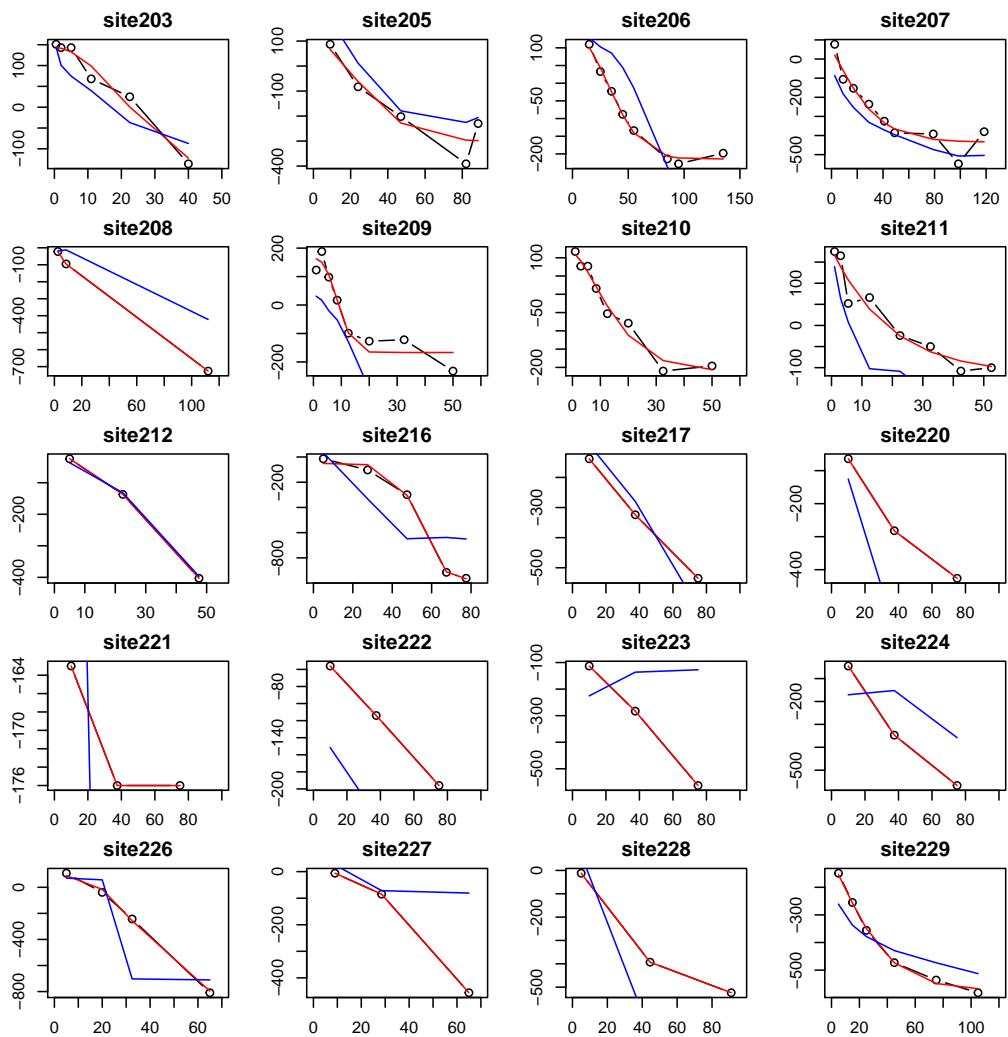
points(x,model(phi = phi.optim,x),col = "red",type ="l")
lines(x,y.simul,col ="blue")
}

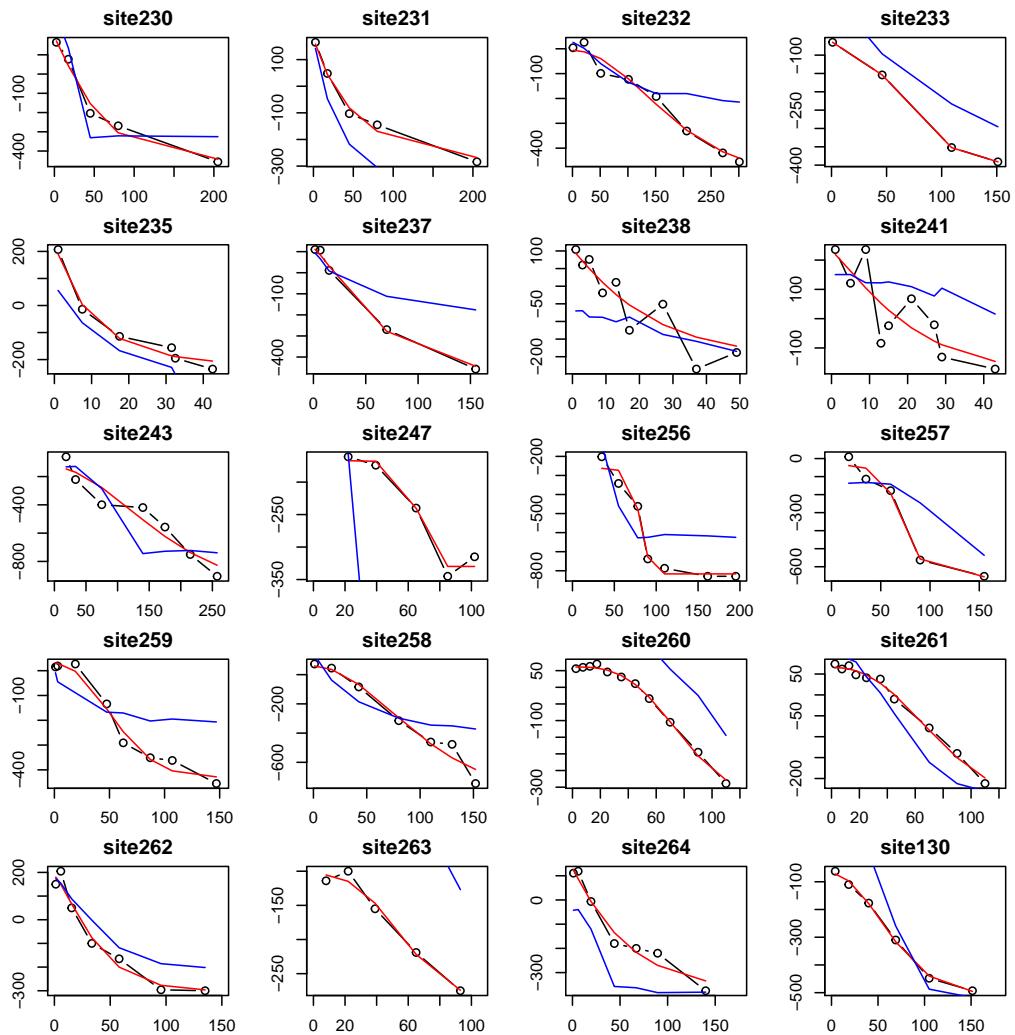
```

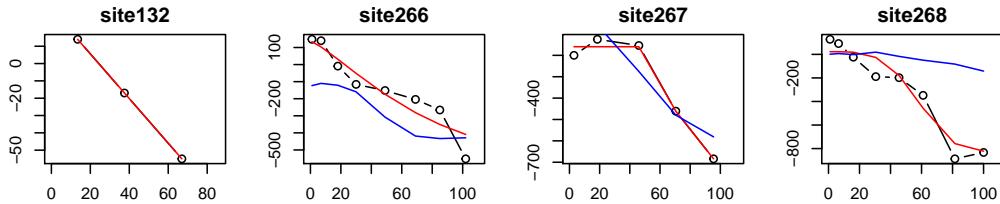












```

# Remarque JAGS ne sait pas traiter les listes

# le nb de mesures par site
size_par_site= sapply(y.sim, function(x){length(unlist(x))})

# le nbr maximale de mesure pour un site
max.length <- max(unlist(size_par_site))

## Pour D14C
## Ajouter Na pour compléter les éléments de la liste
y <- lapply(y.sim, function(v) {c(v, rep(NA, max.length-length(v)))})
## Rbind
y <- do.call(rbind, y)

# Pour la profondeur

```

```

## Ajouter Na pour compléter les éléments de la liste
z<- lapply(z, function(v) { c(v, rep(NA,max.length-length(v))))})
## Rbind
z<-do.call(rbind,z)

```

E Code R :Modèle sous Jags

```

library(R2jags)
library(runjags)
library(coda)

# Apporter le modèle proposé
model.jags <- paste(getwd(),"/carbon-model.txt", sep="")
#file.show(model.jags)

# Data
X = model.mat
RR = matrix(as.numeric(solve(Omega.hat)),ncol =4,nrow = 4)
data = list(v = 5,p = nrow(theta.hat),N = S,y = y,X = X,z = z
,R = Omega.hat,size_par_site = size_par_site,a = 0.001, b= 0.001)

#les valeurs initiales

inits <- list(
  list(theta = theta.hat,Prec = RR,tau = 0.01)
 ,list(theta=matrix(runif(nrow(theta.hat)*4,0.9,1.1)
 ,nrow(theta.hat),4)*theta.hat,
 Prec = diag(1,ncol = 4,nrow = 4),tau = 0.02)
 ,list(theta=matrix(runif(nrow(theta.hat)*4,1,1.1)
 ,nrow(theta.hat),4)*theta.hat,Prec = RR,tau = 0.05)
 )

# Les paramètres à estimer
params <- c("theta","Prec","tau")

# Jags

jagsfit_sim <- jags(data = data,inits = inits,
parameters.to.save = params,n.iter = 75000
, model.file = model.jags,n.chains = 3,n.burnin = 50000,n.thin = 5)

```

F Diagnostic de la convergence

```

library(coda)
# Diagnostic de la convergence avec le package coda
jagsfit.coda <- as.mcmc(jags.fit)

```

```

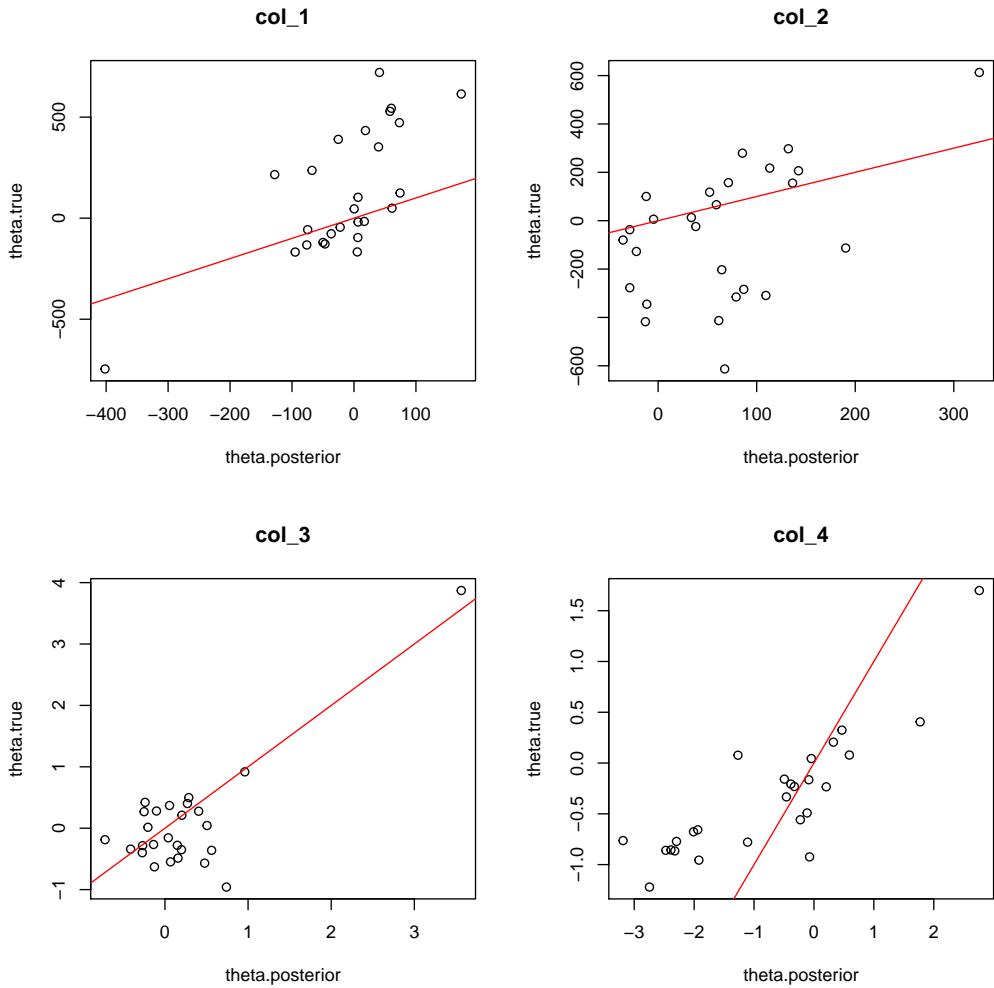
par(mfrow = c(5,4), mar = c(4,4,4,4))
traceplot(jagsfit.coda)
par(mfrow = c(5,4), mar = c(4,4,4,4))
densplot(jagsfit.coda)

# Autocorrelation
par(mfrow= c(5,4), mar = c(4,4,4,4))
autocorr.plot(jagsfit.coda)

# Gelman and Rubin using coda
diag <- gelman.diag(jagsfit.coda,multivariate=FALSE)
no_convergence<- which(diag[1]$psrf[,1] > 1.2)
param_no_convergence<-diag[1]$psrf[no_convergence,1]
length(param_no_convergence)

```

G Test du modèle sous Jags



H Code R : Application sur les données réelles

Le code R ainsi que les bandeaux de prédiction pour les 74 sites d'apprentissage :

```

library(R2jags)
library(runjags)
library(coda)

# Apporter le modèle proposé
model.jags <- paste(getwd(), "/carbon-model.txt", sep="")

```

```

#file.show(model.jags)

# ajouter à la base de donnée une colonne correspondante
#à la profondeur moyenne par couche

Data.selected$mean_level<-0.5*(Data.selected$level_top+Data.selected$level_base)

# Apprentissage et validation
ind_sites=c()
set.seed(123)
for (k in unique(Sol)){
  ind_sites=c(ind_sites,which(Sol==k)[sample(1:length(which(Sol==k)),1)])
}
for (k in unique(Land)){
  ind_sites=c(ind_sites,which(Land==k)[sample(1:length(which(Land==k)),1)])
}
ind_sites=unique(ind_sites)
sites.ap=unique(Data.selected$site_nb)[ind_sites]

# sites à choisir encore pour l'apprentissage :
nbr= 74-length(sites.ap)           ##### à modifier ici !!!!!!!!
se=1:length(unique(Data.selected$site_nb))
ind_restant=sample(se[-(ind_sites)], nbr, replace = F) # choix aléatoire des environ 74 indices de site restant à choisir pour l'apprentissage
sites.ap=c(sites.ap,unique(Data.selected$site_nb)[ind_restant])
ind.ap = c(ind_sites,ind_restant)
sites.validation = unique(Data.selected$site_nb)[-ind.ap]

### Tourner le modèle sur l'ensemble d'apprentissage

# Sélection la base de données correspondante à l'ensemble d'apprentissage
Data.learning = Data.selected[Data.selected$site_nb %in% sites.ap,]
# préparation de y => (D14C) et z =>(la profondeur)
hh = unstack(Data.learning, form=Data.learning$D14C ~ Data.learning$site_nb)
nb_par_site = sapply(hh,function(x){length(unlist(x))})
nb_par_site = as.numeric(nb_par_site)
max.length <- max(nb_par_site)

y <- NULL
z <- NULL
size_par_site <- c()

for (i in 1:length(unique(Data.learning$site_nb))){
  s = Data.learning[Data.learning$site_nb == unique(Data.learning$site_nb)[i]
                 ,c("D14C", "mean_level")]
  y <- rbind(y,c(s$D14C, rep(NA,max.length-nrow(s))))
  z <- rbind(z, c(s$mean_level,rep(NA,max.length-nrow(s))))
  size_par_site[i] <- nrow(s)
}

# Jags
# Data
X = model1.mat[ind.ap,]
RR = matrix(as.numeric(solve(Omega.hat)),ncol =4,nrow = 4)
data = list(v = 5,p = nrow(theta.hat),N = nrow(X),y = y,X = X,
           z = z,R = Omega.hat,size_par_site = size_par_site,

```

```

a = 0.001, b= 0.001)

#les valeurs initiales

inits <- list(
  list(theta = theta.hat,Prec = RR,tau = 0.01),
  list(theta=matrix(runif(nrow(theta.hat)*4,0.9,1.1),
nrow(theta.hat),4)*theta.hat,Prec = diag(1,ncol = 4,nrow =4),
tau = 0.02) ,
  list(theta=matrix(runif(nrow(theta.hat)*4,1,1.1),
nrow(theta.hat), ,4)*theta.hat,Prec = RR,tau = 0.05)
)

# Les paramètres à estimer
params <- c( "theta","Prec","tau")

#jagsfit_real<-jags(data=data,inits=inits,parameters.to.save=params
#,n.iter = 75000,model.file = model.jags,n.chains = #3,n.burnin = 50000
#,n.thin = 5)

#save(jagsfit_real,file="jagsfit_real.Rdata")
load(file = "jagsfit_real.Rdata")

## Diagnostic de la convergence
diag<- jagsfit_real$BUGSoutput$summary[,8]
ind_no_convergence <- which (diag> 1.2)
param_no_convergence<- diag[ind_no_convergence]
length(param_no_convergence)

[1] 0

range(param_no_convergence)

[1] Inf -Inf

#####
# bandeaux de prédiction
#####

n.sims = jagsfit_real$BUGSoutput$n.sims
theta.posterior <- jagsfit_real$BUGSoutput$sims.list$theta
tau.posterior <- jagsfit_real$BUGSoutput$sims.list$tau

par(mfrow = c(5,4), mar = c(2,2,2,2))

for (i in 1:74){# on a 74 sites d'apprentissage

  new.prof = seq(0,300,by=10)
  prof = z[i,1:nb_par_site[i]]
  model.posterior<-matrix(0, nrow = n.sims,ncol=length(new.prof))
  y.posterior <- model.posterior

  for (g in 1:n.sims){


```

```

# variable latente
phi.posterior<-mvrnorm(1,
  model.mat[i, ]%*%jagsfit_real$BUGSoutput$mean$theta,
  solve(jagsfit_real$BUGSoutput$mean$Prec))
# modèle
model.posterior[g, ] <- phi.posterior[1]+phi.posterior[2]*
exp(-(new.prof/exp(phi.posterior[3]))^(exp(phi.posterior[4])))

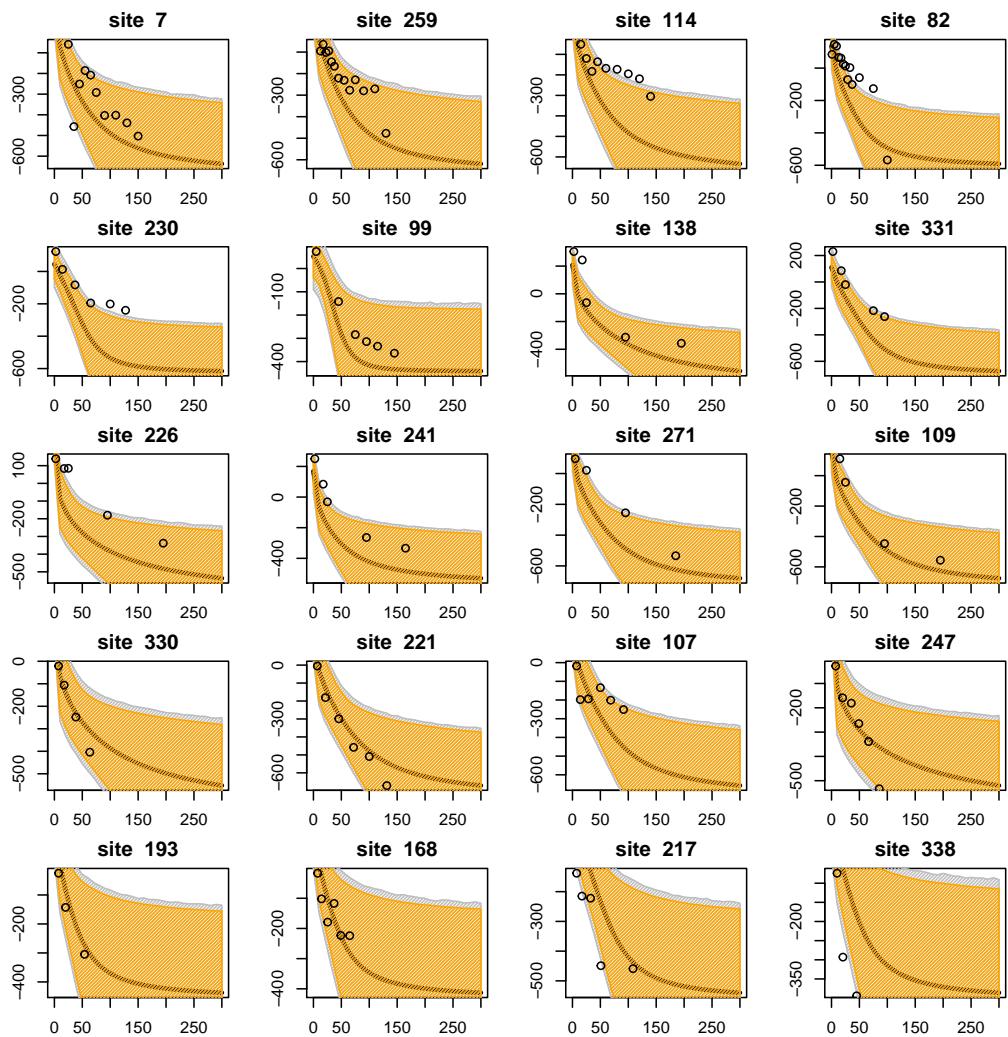
# valeur prédictée
y.posterior[g, ] <- model.posterior[g, ]+rnorm(length(new.prof),
  0, sqrt(1/tau.posterior[g]))
}

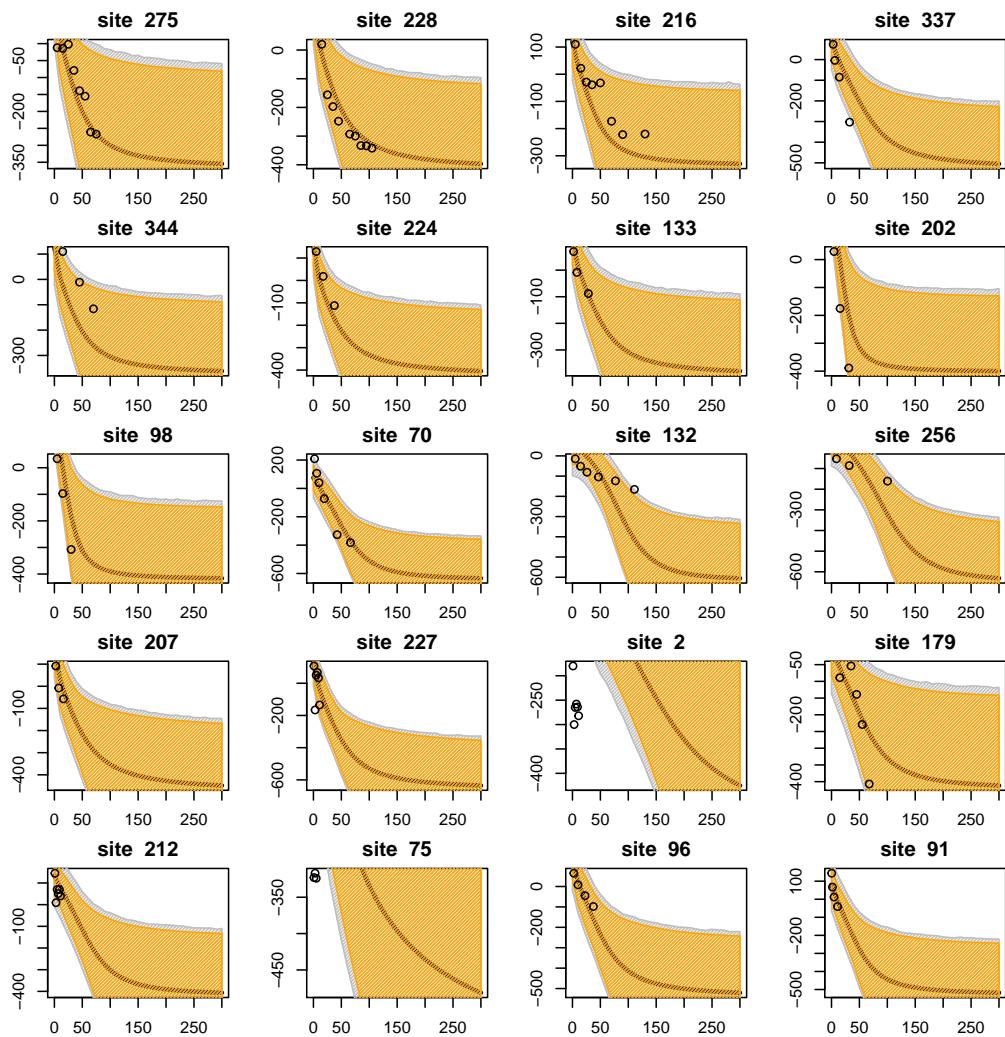
# tracer les bandeaux de prédictions

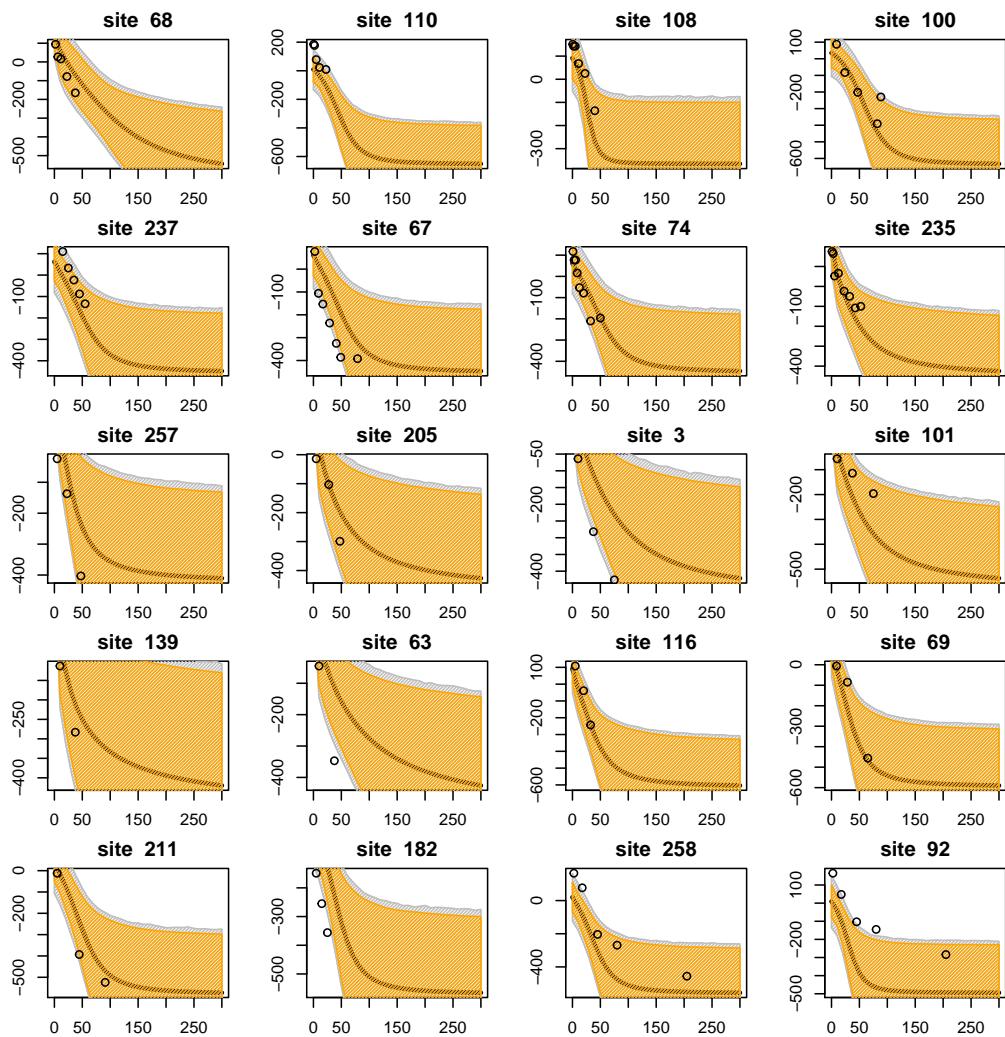
plot(new.prof, apply(model.posterior, 2, mean), typ='l',
  ylab=expression(paste(Delta^14, "C", sep = "")),
  xlab='profondeur', main=paste("site ", sites.ap[i]),
  lwd=3, ylim = c(min(apply(model.posterior, 2, mean)),
  max(y[i, 1:nb_par_site[i]], na.rm = T)))

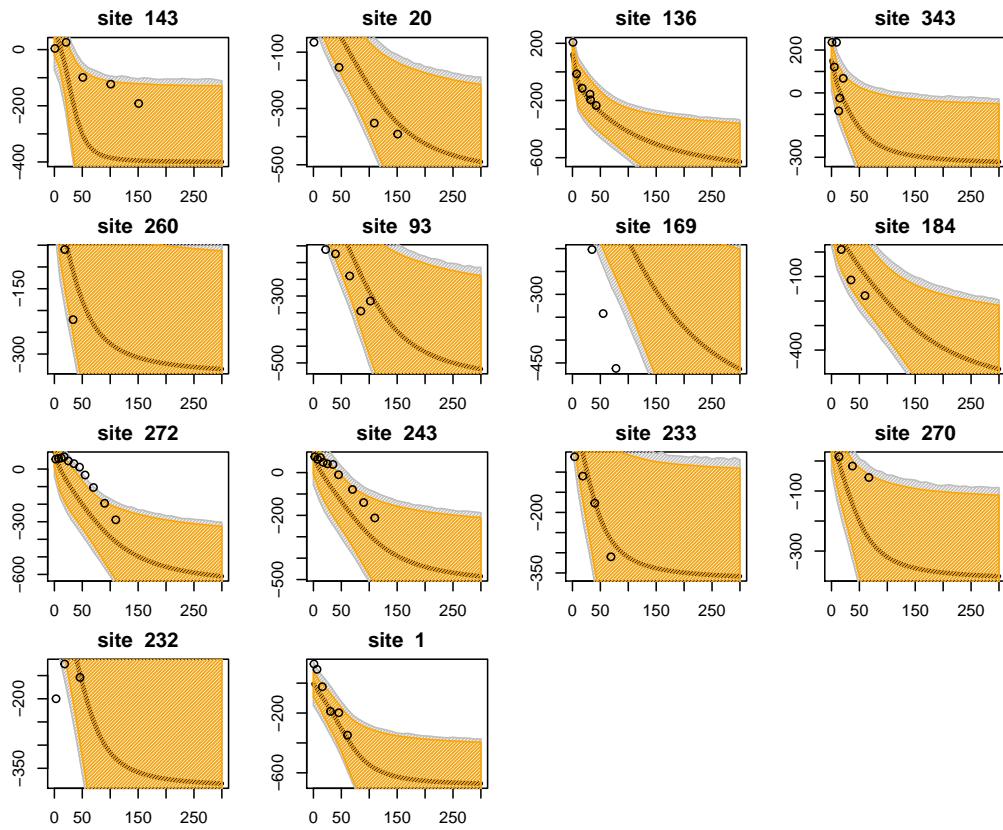
ympolygone=c(apply(y.posterior, 2, quantile, prob=0.05),
  apply(y.posterior, 2, quantile, prob=0.95)[length(new.prof):1])
ypolygone=c(apply(model.posterior, 2, quantile, prob=0.05),
  apply(model.posterior, 2, quantile, prob=0.95)[length(new.prof):1])
xpolygon=c(new.prof, new.prof[length(new.prof):1])
polygon(xpolygon, ympolygone, col="grey", density=c(60))
polygon(xpolygon, ypolygone, col="orange", density=c(60))
points(prof, y[i, 1:nb_par_site[i]])
}

```

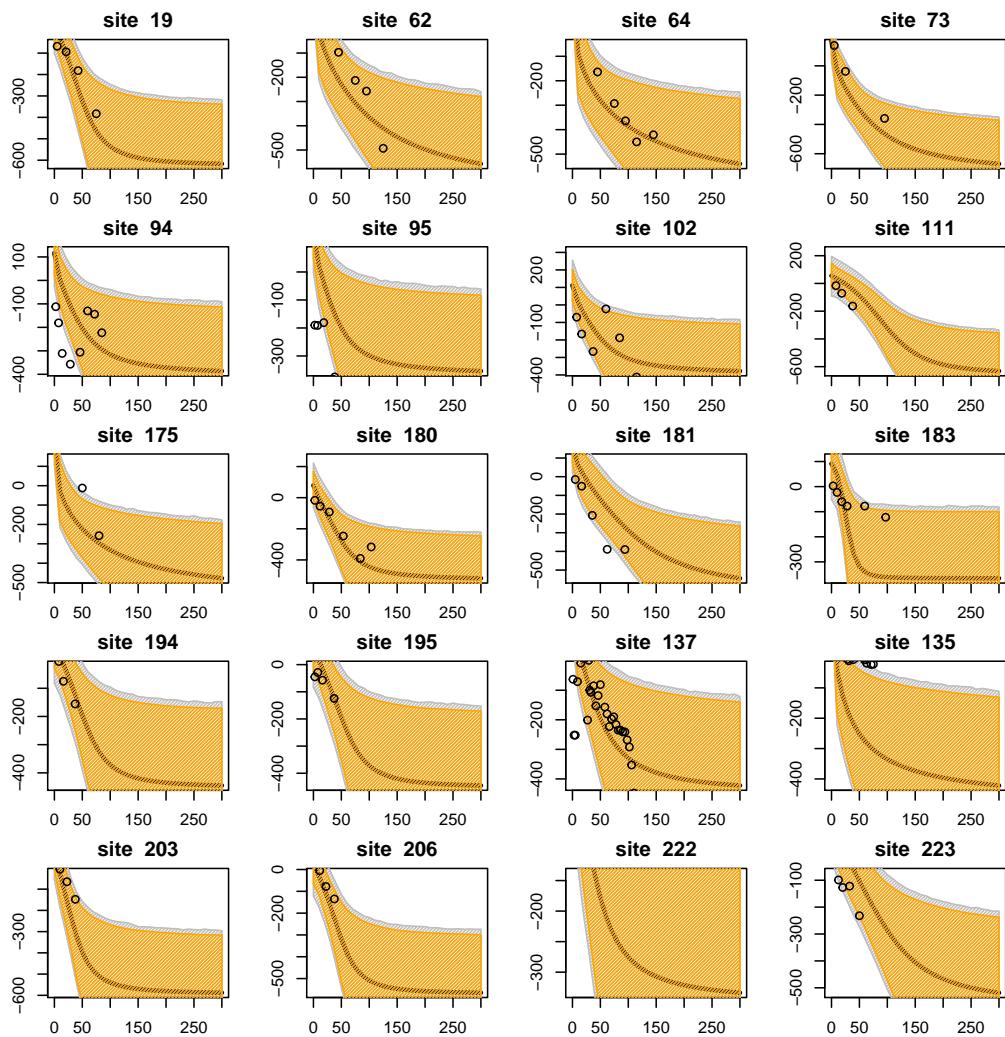


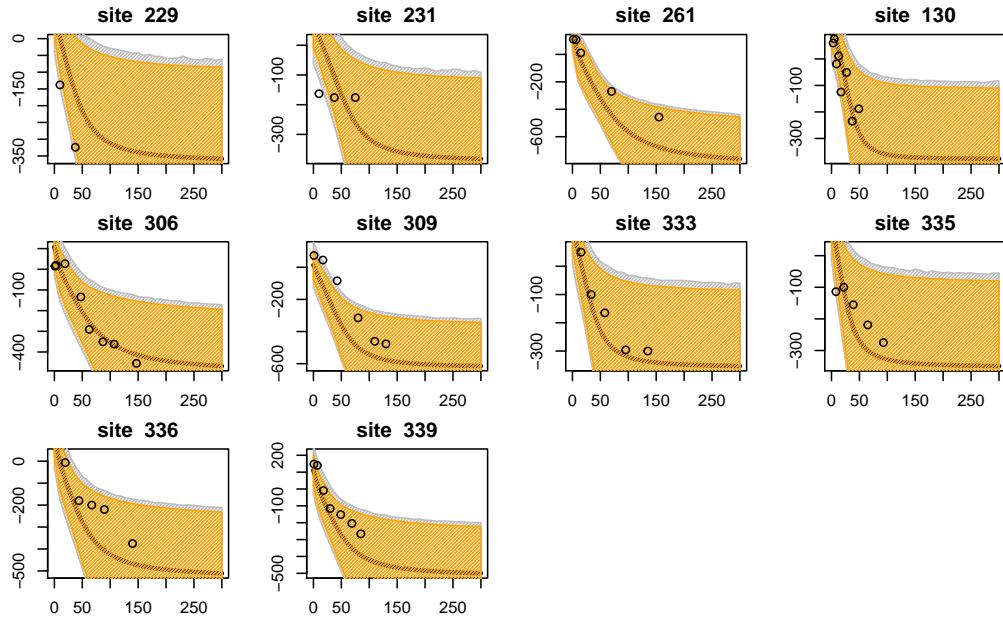






Les bandeaux de prédiction pour les site de validation ainsi que le code r sont donnés ci-dessous :





I Code R : Sélection de variable

```

library(R2jags)
library(coda)

# Apporter le modèle proposé pour faire de la sélection
model.jags <- paste(getwd(),"/carbon-model-selection.txt", sep="")
#file.show(model.jags)

```

```

# préparation de y => (D14C) et z =>(la profondeur)
hh = unstack(Data.selected,form=Data.selected$D14C ~ Data.selected$site_nb)

nb_par_site = sapply(hh,function(x){length(unlist(x))})
nb_par_site = as.numeric(nb_par_site)
max.length <- max(nb_par_site)

y <- NULL
z <- NULL
size_par_site <- c()

for (i in 1:length(unique(Data.selected$site_nb))){
  s = Data.selected[Data.selected$site_nb==
    unique(Data.selected$site_nb)[i],c("D14C","mean_level")]

  y <- rbind(y,c(s$D14C,rep(NA,max.length-nrow(s))))
  z <- rbind(z, c(s$mean_level,rep(NA,max.length-nrow(s))))
  size_par_site[i] <- nrow(s)
}

# Data
X = model.mat
RR = matrix(as.numeric(solve(Omega.hat)),ncol =4,nrow = 4)
data=list(N=S,y=y,X=X,z=z,R=Omega.hat,
          size_par_site= size_par_site, pos=c(1,rep(2,7),
          rep(3,10),4:11),ncov=26)

#les valeurs initiales

inits <- list(
  list(Prec = RR,phi = phi.sim, sd_y = 20,beta = theta.hat,
       ind1 = sample(1:2 ,replace = T ,size = 26),
       ind2 = sample(1:2 ,replace = T ,size = 26),
       ind3 = sample(1:2 ,replace = T ,size = 26),
       ind4= sample(1:2 ,replace = T ,size = 26) ),
  list(Prec = RR,phi = matrix(runif(nrow(phi.sim)*4,0.9,1.1),
      nrow(phi.sim),4)*phi.sim, sd_y = 11,
      beta=matrix(runif(nrow(theta.hat)*4,0.9,1.1),
      nrow(theta.hat),4)*theta.hat,
      ind1 = sample(1:2 ,replace = T ,size = 26),
      ind2 = sample(1:2 ,replace = T ,size = 26),
      ind3 = sample(1:2 ,replace = T ,size = 26),
      ind4 = sample(1:2 ,replace = T ,size = 26)),
  ,list(Prec=RR,phi=matrix(runif(nrow(phi.sim)*4,1,1.1),
      nrow(phi.sim),4)*phi.sim,
      sd_y = 10,beta = theta.hat,
      ind1 = sample(1:2 ,replace = T ,size = 26),
      ind2 = sample(1:2 ,replace = T ,size = 26),
      ind3 = sample(1:2 ,replace = T ,size = 26),
      ind4 = sample(1:2 ,replace = T ,size = 26)))
)

# Les paramètres à estimer
params <- c("beta","Prec","sd_y","gamma1","gamma2","gamma3","gamma4")

```

```

# Jags
#jagsfit_selection <- jags(data = data, inits =
#                           parameters.to.save = params, n.iter = 75000,
#                           model.file = model.jags, n.chains = 3,
#                           n.burnin = 50000, n.thin = 5)

#save(jagsfit_selection, file="jagsfit-selection.Rdata")
#load(file = "jagsfit-selection.Rdata")

y.posterior <- list()
for (i in 1:104){# on a 104 sites
  prof = z[i,1:size_par_site[i]]

  phi.posterior <- mvrnorm(1,
    model.mat[i,] %*% jagsfit_selection$BUGSoutput$mean$beta,
    solve(jagsfit_selection$BUGSoutput$mean$Prec))

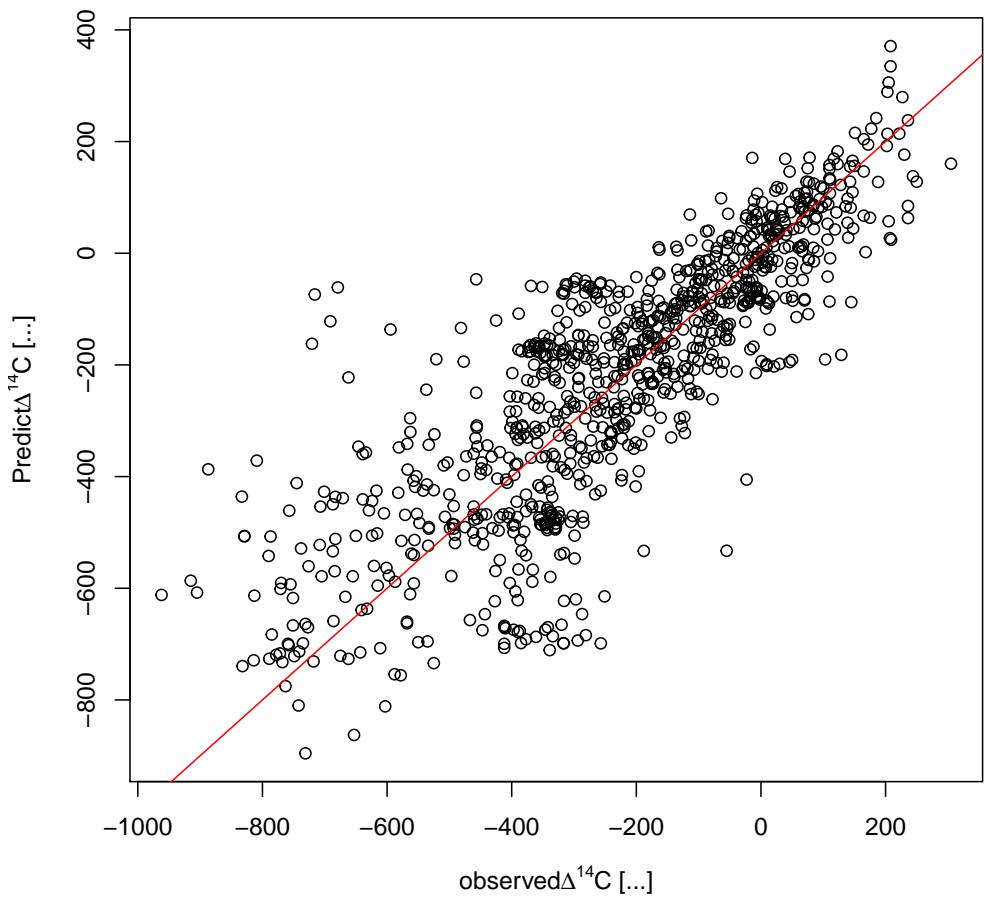
  model.posterior <- phi.posterior[1]+phi.posterior[2]*
    exp(-(prof/exp(phi.posterior[3]))^(exp(phi.posterior[4])))

  y.posterior[[i]]<-model.posterior+
    rnorm(size_par_site[i],0,
      sqrt(jagsfit_selection$BUGSoutput$mean$sd_y))
}

plot(Data.selected$D14C,unlist(y.posterior), type = "p",
  xlab = expression(paste("observed", Delta ^{14} ,C , " [%]" )),
  ylab = expression(paste("Predict", Delta ^{14} ,C , " [%]" )))

abline(a=0,b=1, col = "red")

```



Références

- [1] Edward I.George ; Robert E.McCulloch. Variable selection via gibbs sampling. *Jounal of the American Statistical Association*, Vol.88, No.423,881-889, Sep.,1993.
- [2] Jerome Balesdent Jordane Mathieu, Hatté Christine and Eric Parent. Deep soil carbon dynmic are driven more by soil type than by climate : A worldwide meta-analysis of radiocarbon profiles. *Global Change Biology*, June 2015.