Analisi gerarchica bayesiana dei

Profili verticali di nutrienti del suolo

SOC, N e P in 40 campi agricoli in Tanzania

Bortoletto Davide Faedo Piero Paganelli Gabriele Zannini Pietro

Statistica Iterazione · Università degli Studi di Padova, a.a. 2025/26

8 giugno 2026

Abstract

La distribuzione verticale dei nutrienti del suolo è un indicatore chiave della qualità agronomica, ma i meccanismi che determinano le differenze tra campi nella forma del profilo di profondità restano poco quantificati. In questo lavoro analizziamo i profili verticali di carbonio organico (SOC), azoto totale e fosforo totale in 40 campi agricoli, con misurazioni fino a sei profondità per campo, per un totale di 220 osservazioni. Sviluppiamo un modello bayesiano gerarchico con pendenza proporzionale che stima esplicitamente come la variabilità inter-campo del profilo verticale cambi con la profondità. I risultati mostrano che i campi con maggiore contenuto medio di carbonio organico presentano profili significativamente più uniformi: la fertilità carboniosa si conserva meglio in profondità. Per l'azoto questo *pattern* è assente; per il fosforo è presente con evidenza moderata. La selezione predittiva identifica la tessitura fine (contrasto frazioni fini/sabbia) come principale predittore *within-field*, mentre le pratiche di gestione risultano essenzialmente non significative dopo aver controllato per la struttura fisica del suolo. A differenza degli approcci standard basati su confronti stratificati per profondità, il modello proposto fornisce stime inferenziali dirette della relazione tra livello medio di campo e forma del profilo verticale, aprendo la strada a una caratterizzazione quantitativa dell'eterogeneità agronomica tra campi. Il codice è disponibile su: `https://github.com/GabriPaganelli/crop-soil-analysis`.

Indice

1 Introduzione **3**

2 Dati **3**

3 Analisi esplorativa **5**

4 Modelli statistici e inferenza **6**
4.1 Analisi preliminari 6
4.2 Struttura del modello finale 7
4.3 *Prior* 8
4.4 MCMC e diagnostica 8

5 Risultati e validazione del modello **9**
5.1 Effetti fissi 9
5.2 Parametri strutturali 9
5.3 Correlazioni tra nutrienti a livello di campo 10
5.4 Selezione delle covariate 10
5.5 Confronto tra strutture gerarchiche 10
5.6 Analisi di robustezza 11

6 Discussione dei risultati **12**
6.1 Evidenze sulla struttura di correlazione 12
6.2 Evidenze sulle covariate fisiche 13
6.3 Evidenze sugli effetti gestionali 13

7 Conclusioni **14**

A Statistiche descrittive delle variabili nel *dataset* **16**

B *Trace plot* e *posterior predictive check* **16**

1. Introduzione

La distribuzione verticale dei nutrienti del suolo è un indicatore fondamentale della qualità agronomica: un campo che conserva materia organica anche negli strati profondi risponde in modo diverso agli input gestionali rispetto a uno in cui la fertilità è concentrata nella sola superficie. Quantificare questa variabilità inter-campo, intesa non soltanto come differenza nel livello medio ma anche nella forma del profilo verticale, è un passaggio necessario per caratterizzare l'eterogeneità agronomica del territorio.

Gli approcci standard modellano i profili con ANOVA stratificata per profondità o con modelli misti a sola intercetta casuale, analizzando i nutrienti separatamente. Questi metodi stimano il livello medio di campo ma non la variabilità nella pendenza verticale, né la covariazione tra nutrienti a livello di campo. Con poche decine di campi ciascuno misurato a più profondità, la stima di queste strutture richiede una regolarizzazione esplicita, che l'approccio bayesiano fornisce in modo naturale attraverso le distribuzioni a priori. In questo lavoro si adotta un modello gerarchico multivariato che stima entrambe le strutture, con l'obiettivo di valutare se e in che misura le proprietà fisiche del suolo e le pratiche di gestione ne spiegano la variabilità.

2. Dati

Il *dataset* comprende 40 campi agricoli nel sud della Tanzania. Per ognuno di essi sono state campionate cinque o sei profondità (20, 30, 40, 50, 60 e, in alcuni casi, 80 cm), per un totale di 220 osservazioni. Ogni carotaggio è stato suddiviso in sezioni verticali e per ciascuna l'analisi chimica è stata condotta su un campione di terra omogeneizzata, così da ottenere una misura integrata sull'intero segmento verticale piuttosto che su un singolo punto.

Le tre variabili risposta (carbonio organico, azoto totale e fosforo totale) sono espresse come percentuale in peso del suolo secco. Nella scala originale le distribuzioni sono marcatamente asimmetriche e il decadimento con la profondità è compatibile con un andamento log-lineare, come riportato in letteratura per i profili di nutrienti nei suoli tropicali. Si è pertanto adottata la trasformazione logaritmica:

$$y_r = \log(\mathrm{Perc}_r), \quad r \in \{\mathrm{SOC}, N, P\},$$

che stabilizza la varianza e linearizza la relazione con la profondità. Le tre percentuali non condividono alcun vincolo di somma e i valori osservati rimangono ampiamente al di sotto del 100%, rendendo superfluo qualsiasi aggiustamento composizionale.

Alcune caratteristiche del suolo differiscono all'interno dello stesso campo con la profondità. Il pH e la densità apparente (`BulkDensity`) descrivono l'acidità e la compattazione del suolo lungo il profilo. La composizione granulometrica richiede invece un trattamento particolare: argilla, limo e sabbia sommano per costruzione al 100%, definendo un sistema composizionale in cui le tre frazioni non sono linearmente indipendenti. Inserirle direttamente come predittori introdurrebbe una dipendenza lineare perfetta e correlazioni spurie tra i coefficienti. Si è pertanto applicata la trasformazione Isometric Log-Ratio (ILR) (Egozcue et al., 2003), che ricava due coordinate ortogonali prive di vincolo:

$$T_1 = \sqrt{\tfrac{1}{2}}\,\log\!\left(\frac{\mathrm{Clay}}{\mathrm{Silt}}\right), \qquad T_2 = \sqrt{\tfrac{2}{3}}\,\log\!\left(\frac{\sqrt{\mathrm{Clay}\cdot\mathrm{Silt}}}{\mathrm{Sand}}\right).$$

Le due coordinate, denominate `Texture1` e `Texture2`, hanno un'interpretazione diretta: T_1 cattura il contrasto tra argilla e limo all'interno della frazione fine, T_2 il gradiente tra le frazioni fini nel loro insieme e la sabbia. La distribuzione delle osservazioni nello spazio tessiturale è illustrata in Figura 1: le 220 osservazioni ricadono in sei classi del triangolo tessiturale USDA (United States Department of Agriculture), con una concentrazione nelle zone limosa e franco-limosa.

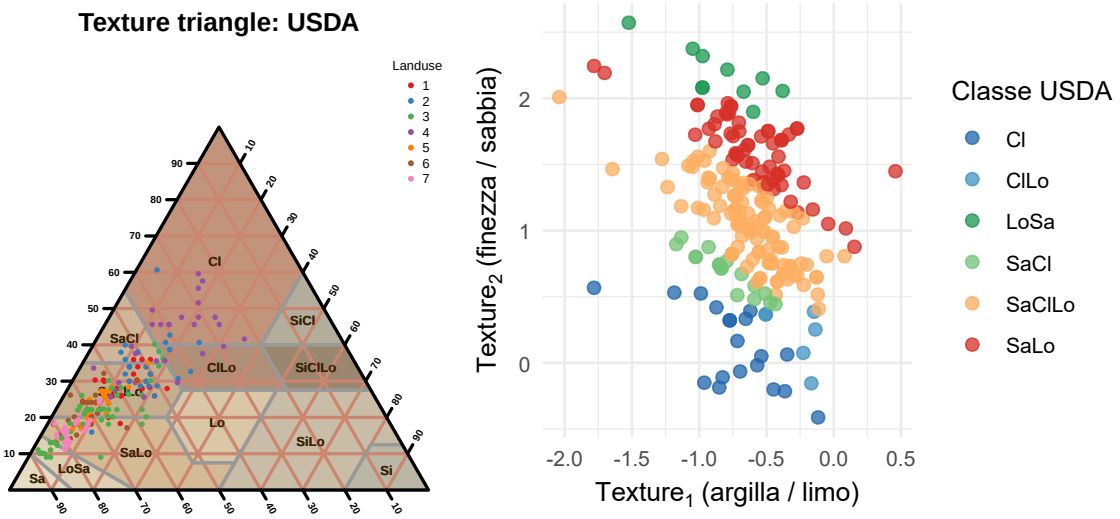


Figura 1: Tessitura del suolo: triangolo tessiturale colorato per categoria d'uso del suolo (sinistra) e coordinate ILR colorate per classe tessiturale (destra).

Le caratteristiche di gestione sono fisse per campo. `OnFarm` indica se il campo è collocato nell'area principale della fattoria, `Irrigate` e `Fertilised` registrano il regime idrico e di fertilizzazione, `N_Natural` individua i cinque campi non coltivati inclusi come riferimento naturale. Queste quattro variabili presentano una struttura di co-occorrenza parzialmente deterministica: come mostra la Tabella 1, nessun campo *on-farm* è privo di fertilizzazione. Il *dataset* include anche la variabile `Landuse`, che codifica le sette combinazioni di gestione osservate; si è scelto di inserire i quattro effetti individualmente, così da preservare la leggibilità dei singoli coefficienti e mantenere la possibilità di separare, almeno parzialmente, il contributo delle singole pratiche.

	OnFarm 0	**OnFarm 1**	**Totale**
Fertilised 0	20	0	20
Fertilised 1	10	10	20
Totale	30	10	40

Tabella 1: Distribuzione di `OnFarm` e `Fertilised` per i 40 campi.

La Figura 2 mostra la distribuzione spaziale dei 40 campi: essi non coprono uniformemente il territorio, bensì si concentrano in due aree ben distinte all'interno di una regione compatta di circa 20 km nel sud della Tanzania. Il gruppo settentrionale comprende prevalentemente campi *off-farm* con diversi regimi di fertilizzazione, mentre il gruppo meridionale include la maggior parte dei campi *on-farm*, tutti fertilizzati. Questa sovrapposizione tra prossimità geografica e pratiche di gestione costituisce un confondimento strutturale del disegno di

campionamento: non è possibile separare con certezza l'effetto della posizione spaziale da quello delle variabili di gestione. Le coordinate latitudine e longitudine, sebbene disponibili, non sono state quindi incluse come predittori nel modello, in quanto una loro inclusione implicherebbe un andamento lineare con la posizione geografica privo di interpretazione agronomica e difficilmente generalizzabile.

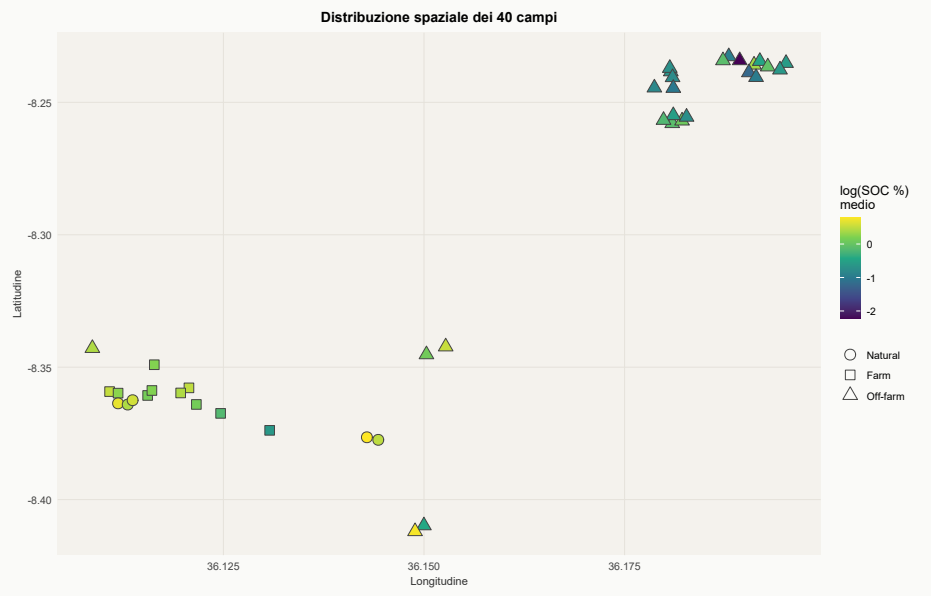


Figura 2: Distribuzione spaziale dei 40 campi nel sud della Tanzania. Il gruppo settentrionale comprende prevalentemente campi *off-farm*; quello meridionale i campi *on-farm*, tutti fertilizzati. La sovrapposizione geografica con le pratiche di gestione costituisce un confondimento strutturale del disegno.

3. Analisi esplorativa

La Figura 3 mostra i profili verticali dei 40 campi per le tre risposte in scala logaritmica. Il decadimento con la profondità è evidente per carbonio e azoto (pendenza mediana -1.30 e -0.72 in scala log-log; solo 3 campi su 40 mostrano un profilo crescente per ciascuna risposta), mentre per il fosforo la tendenza è molto più debole (13 campi su 40 con profilo crescente). Si osserva una notevole variabilità tra le traiettorie, che giustifica l'introduzione di un'intercetta casuale per campo. Inoltre, le traiettorie individuali non sono parallele: per il carbonio in particolare, i campi con concentrazioni più elevate tendono ad avere profili più piatti, conservando il carbonio anche negli strati profondi; per azoto e fosforo questo *pattern* non è visibile. Ciò indica che la velocità di decadimento con la profondità varia sistematicamente da campo a campo, motivando l'inclusione di una pendenza casuale.

Considerando le medie non condizionate per campo, carbonio e azoto mostrano una correlazione positiva forte ($r \approx 0.73$): i campi più ricchi in carbonio tendono ad esserlo anche in azoto. Il fosforo mostra correlazioni più contenute con gli altri due nutrienti ($r \approx 0.49$), indicando una parziale autonomia del suo ciclo.

Per carbonio e azoto si osserva un'apparente risalita dei valori all'ultima profondità di campionamento (80 cm). Questo è tuttavia riconducibile a un artefatto nel disegno di campionamento: non tutti i campi raggiungono 80 cm di profilo campionato, e le traiettorie visibili a tale quota rappresentano pertanto un sottoinsieme selezionato di campi. Considerando

esclusivamente i campi con profilo completo fino a 80 cm, si osserva che il decadimento rimane monotono anche all'ultima profondità per carbonio e azoto.

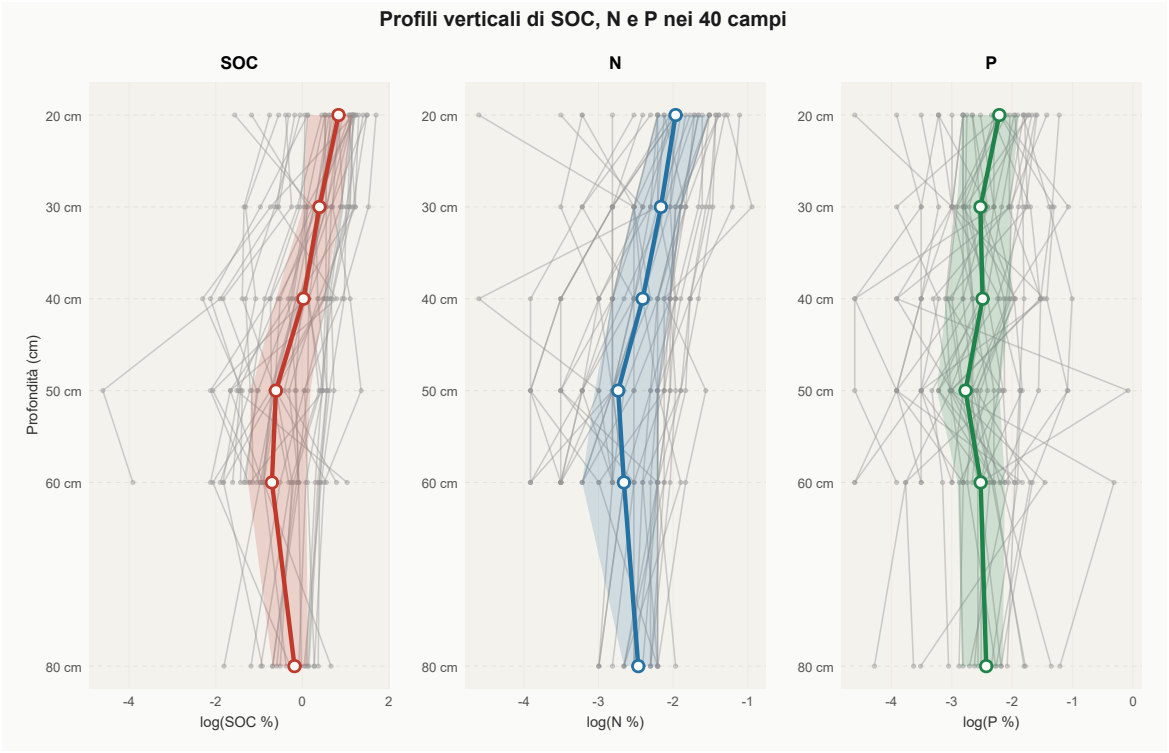


Figura 3: Profili verticali di SOC, N e P nei 40 campi (scala log). Linee grigie: singoli campi; linea colorata: mediana globale per profondità; banda: intervallo interquartile.

4. Modelli statistici e inferenza

4.1. Analisi preliminari

Come punto di partenza, le tre risposte sono state analizzate separatamente tramite modelli lineari misti con intercetta casuale per campo. I predittori considerati sono le covariate fisiche (profondità, Texture_1, Texture_2, densità apparente e pH) e le variabili di gestione: posizione in fattoria, regime idrico, fertilizzazione e appartenenza a classe naturale di riferimento. Le covariate continue (profondità in scala logaritmica, densità apparente e pH) sono state standardizzate prima della stima, in modo da rendere i coefficienti confrontabili tra loro; Texture_1 e Texture_2, essendo già nella loro scala naturale di coordinate ILR, non richiedono ulteriore trasformazione.

Per la profondità, tra le diverse forme funzionali testate (lineare, quadratica, logaritmica), la relazione log-log, corrispondente a una legge di potenza, risulta la più parsimoniosa e ha un'interpretazione diretta nel contesto dei profili verticali di nutrienti: un incremento percentuale fisso della profondità corrisponde a una variazione costante della risposta in scala logaritmica.

Per carbonio e azoto emerge un quadro coerente: il decadimento con la profondità è il predittore dominante; Texture_2 e densità apparente mostrano effetti negativi stabili, mentre pH e Texture_1 non aggiungono potere predittivo. Le variabili di gestione non mostrano segnali apprezzabili per queste due risposte. Per il fosforo il comportamento è opposto: la profondità non contribuisce in modo sistematico, mentre la posizione in fattoria e l'irrigazione mostrano

un'associazione positiva con le concentrazioni. L'analisi frequentista rileva inoltre una lieve autocorrelazione verticale nei profili del carbonio, assente per azoto e fosforo, suggerendo che la dipendenza verticale del carbonio meriti attenzione nella specificazione del modello finale.

I coefficienti di correlazione intraclasse indicano che una quota rilevante della varianza residua è stabile a livello di campo, confermando la necessità di una struttura gerarchica. Tuttavia, modellare le tre risposte indipendentemente nasconde un'informazione importante: le intercette casuali di carbonio e azoto, ovvero il livello medio di ciascun campo al netto di profondità e covariate, mostrano una correlazione positiva (Figura 4), indicando che i campi con più carbonio tendono ad avere anche più azoto, una dipendenza che modelli univariati non possono catturare.

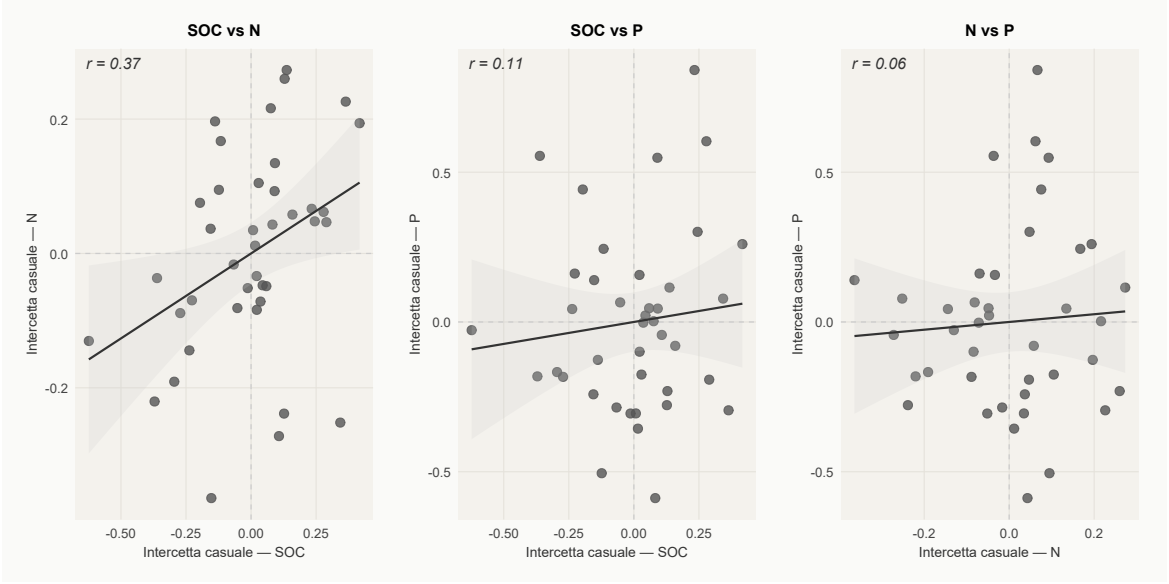


Figura 4: Intercette casuali di campo stimate separatamente per le tre risposte.

Questo quadro motiva l'adozione di un approccio bayesiano multivariato. Con 40 campi, stimare in modo affidabile la struttura di correlazione tra effetti casuali richiede una regolarizzazione esplicita, che le *prior* debolmente informative forniscono in modo trasparente e verificabile. L'approccio bayesiano offre inoltre distribuzioni a posteriori complete per tutti i parametri, incluse le correlazioni tra nutrienti, quantità di interesse primario, nonché una quantificazione dell'incertezza coerente con la struttura gerarchica del problema.

4.2. Struttura del modello finale

Per ogni risposta $r \in \{1, 2, 3\}$ (carbonio organico, azoto e fosforo) e osservazione i nel campo j:

$$y_{r,ij} \sim \mathcal{N}(\mu_{r,ij}, \sigma_r^2),$$

$$\mu_{r,ij} = \alpha_r + u_{rj} + v_{rj} \log B_i + \boldsymbol{\gamma}_r^\top \mathbf{x}_{W,i} + \boldsymbol{\beta}_r^\top \mathbf{x}_{B,j}.$$

Il vettore $\mathbf{x}_{W,i}$ raccoglie le cinque covariate *within-field* (profondità logaritmica, Texture_1, Texture_2, densità apparente, pH); $\mathbf{x}_{B,j}$ le quattro variabili di gestione. I termini u_{rj} e v_{rj} sono rispettivamente l'intercetta e la pendenza casuale di campo per la risposta r: il primo

cattura il livello medio del campo, il secondo consente alla velocità di decadimento con la profondità di variare tra campi.

I sei effetti casuali per campo sono modellati congiuntamente come

$$\mathbf{b}_j = (u_{1j}, v_{1j}, u_{2j}, v_{2j}, u_{3j}, v_{3j})^\top \sim \mathcal{N}_6(\mathbf{0}, \Sigma),$$

con

$$\Sigma = \mathrm{diag}(\boldsymbol{\tau})\, \Omega\, \mathrm{diag}(\boldsymbol{\tau}), \qquad \Omega \sim \mathrm{LKJ}(2).$$

La matrice Ω è 6×6 e descrive sia il legame intercetta-pendenza all'interno di ciascuna risposta, sia le correlazioni tra nutrienti a livello di campo. Queste ultime catturano la covariazione biologica tra nutrienti, abilitando un'inferenza più coerente sugli effetti di gestione.

4.3. *Prior*

Tabella 2: *Prior* del modello finale.

Parametro	Dimensione	*Prior*
α_r	scalare, per $r = 1, 2, 3$	$\mathcal{N}(0,\ 2)$
$\mathbf{z}_j$	$\mathbb{R}^6$, per $j = 1, \ldots, J$	$\mathcal{N}_6(\mathbf{0}, \mathbf{I}_6)$
Ω	matrice di correlazione 6×6	LKJ(2)
$\boldsymbol{\tau}$	$\mathbb{R}^6_+$	$\mathcal{N}^+(0,\ 0.5)$
$\boldsymbol{\gamma}_r$	$\mathbb{R}^{K_W}$, per $r = 1, 2, 3$	$\mathcal{N}(\mathbf{0}, \mathbf{I})$
$\boldsymbol{\beta}_r$	$\mathbb{R}^{K_B}$, per $r = 1, 2, 3$	$\mathcal{N}(\mathbf{0}, \mathbf{I})$
σ_r	scalare positivo, per $r = 1, 2, 3$	$\mathcal{N}^+(0,\ 0.5)$
Totale	$6J + 54 = 294$	

I vettori $\mathbf{z}_j$ corrispondono alla parametrizzazione non centrata degli effetti casuali: invece di campionare $\mathbf{b}_j$ direttamente, si campionano variabili ausiliarie standard da cui $\mathbf{b}_j$ è ricavato deterministicamente, migliorando l'efficienza del campionamento MCMC nelle strutture gerarchiche.

Le *prior* adottate sono debolmente informative: sulla scala standardizzata delle covariate, le normali centrate in zero introducono una regolarizzazione simmetrica dei coefficienti, le *half-normal* sui parametri di scala favoriscono valori contenuti senza escludere quelli elevati se supportati dai dati, e la LKJ(2) regolarizza lievemente le correlazioni verso zero lasciando spazio a strutture sostanziali.

4.4. MCMC e diagnostica

L'inferenza è stata condotta tramite Stan (interfaccia R `cmdstanr`), con quattro catene indipendenti, 3000 iterazioni di *warm-up* e 5000 di campionamento per catena. Per accomodare la complessità della struttura a effetti casuali 6D sono stati impostati `adapt_delta` $= 0.97$ e `max_treedepth` $= 11$.

La convergenza è risultata ottima: nessuna divergenza, $\hat{R} < 1.01$ per tutti i parametri monitorati e dimensioni effettive del campione adeguate. *Trace plot* e *posterior predictive check* sono riportati in Appendice B.

5. Risultati e validazione del modello

Le stime sono riportate come mediane a posteriori con intervalli di credibilità al 90%; i parametri la cui stima esclude chiaramente lo zero sono evidenziati in grassetto nelle tabelle.

5.1. Effetti fissi

Tabella 3: Stime a posteriori degli effetti fissi (mediana e CI al 90%).

Covariata	SOC	N	P
Within-field ($\boldsymbol{\gamma}_r$)			
`logBottom`	**−0.366** [−0.464, −0.270]	**−0.210** [−0.272, −0.148]	−0.075 [−0.160, 0.010]
`Texture1` (ILR_1)	−0.060 [−0.143, 0.024]	−0.024 [−0.078, 0.030]	0.004 [−0.079, 0.088]
`Texture2` (ILR_2)	**−0.503** [−0.640, −0.361]	**−0.284** [−0.371, −0.195]	−0.068 [−0.213, 0.077]
`BulkDensity`	**−0.237** [−0.380, −0.085]	**−0.179** [−0.270, −0.084]	−0.122 [−0.278, 0.033]
`pH`	−0.032 [−0.122, 0.055]	0.003 [−0.052, 0.058]	0.016 [−0.074, 0.105]
Between-field (gestione) ($\boldsymbol{\beta}_r$)			
`OnFarm`	−0.137 [−0.528, 0.274]	**+0.294** [0.037, 0.567]	**+0.684** [0.239, 1.129]
`Irrigate`	**+0.260** [0.022, 0.496]	−0.146 [−0.303, 0.009]	**+0.404** [0.121, 0.683]
`Fertilised`	+0.011 [−0.293, 0.305]	−0.116 [−0.316, 0.074]	−0.346 [−0.683, 0.002]
`N_Natural`	+0.012 [−0.392, 0.402]	+0.176 [−0.108, 0.447]	−0.096 [−0.555, 0.365]

La Tabella 3 riassume le stime degli effetti fissi. Il *pattern* dominante è un'asimmetria sistematica tra carbonio e azoto da un lato, e fosforo dall'altro.

Per carbonio e azoto, la profondità in scala logaritmica e la tessitura fine ($\mathrm{Texture}_2$, contrasto tra frazioni fini e sabbia) sono i predittori *within-field* più rilevanti: all'aumentare della profondità e del contenuto relativo di sabbia, le concentrazioni diminuiscono con un'intensità comparabile per entrambi i nutrienti. La densità apparente contribuisce in misura più contenuta, con lo stesso segno negativo. pH e contrasto argilla/limo ($\mathrm{Texture}_1$) non mostrano effetti apprezzabili per nessuna risposta.

Per il fosforo, nessuna covariata fisica *within-field* produce un intervallo di credibilità lontano da zero. Sono invece le pratiche di gestione a mostrare segnali più chiari: campi *on-farm* e irrigati tendono ad avere concentrazioni di fosforo più elevate. Va tuttavia osservato che la gestione è correlata con la posizione geografica dei campi, cosicché la stima degli effetti gestionali incorpora potenzialmente confondimento spaziale non osservato.

5.2. Parametri strutturali

Tabella 4: Stime a posteriori dei parametri strutturali (mediana e CI al 90%).

Parametro	SOC	N	P
α_r	−0.19 [−0.50, 0.14]	**−2.56** [−2.77, −2.34]	**−2.62** [−2.99, −2.25]
$\tau_{\alpha,r}$	**0.327** [0.234, 0.436]	**0.211** [0.150, 0.287]	**0.396** [0.297, 0.520]
$\tau_{\beta,r}$	**0.215** [0.106, 0.317]	**0.148** [0.080, 0.214]	**0.085** [0.009, 0.185]
ρ_r	**0.519** [0.119, 0.795]	−0.148 [−0.495, 0.235]	0.295 [−0.291, 0.705]
σ_r	**0.524** [0.474, 0.581]	**0.316** [0.286, 0.351]	**0.551** [0.504, 0.605]

La Tabella 4 riporta le stime dei parametri che descrivono la variabilità tra campi. Tutti e tre i nutrienti mostrano una dispersione sostanziale nelle intercette casuali di campo ($\tau_{\alpha,r}$): i campi differiscono nel loro livello medio indipendentemente dalle covariate osservate. La

variabilità delle pendenze ($\tau_{\beta,r}$) è chiaramente diversa da zero per carbonio e azoto, mentre per il fosforo l'intervallo di credibilità sfiora lo zero, indicando profili più uniformi tra campi.

Il risultato più rilevante riguarda la correlazione tra intercetta e pendenza casuale del carbonio (ρ_{SOC}): essa è positiva e l'intervallo di credibilità al 90% esclude lo zero. I campi con un livello medio di carbonio più alto tendono ad avere profili verticali più piatti, conservando il carbonio anche negli strati profondi. Per l'azoto questa correlazione è debole e compatibile con zero; per il fosforo l'incertezza è troppo elevata per trarre conclusioni.

5.3. Correlazioni tra nutrienti a livello di campo

Tabella 5: Correlazioni cross-risposta a livello di campo (mediana e CI 90%).

Coppia	Correlazione intercette	Correlazione pendenze
SOC – N	$+\mathbf{0.386}$ [0.038, 0.667]	+0.289 [−0.152, 0.655]
SOC – P	+0.112 [−0.219, 0.428]	+0.114 [−0.442, 0.589]
N – P	+0.052 [−0.283, 0.376]	−0.139 [−0.603, 0.410]

La Tabella 5 riporta le correlazioni tra gli effetti casuali dei tre nutrienti a livello di campo. L'unica correlazione che emerge con chiarezza è quella tra le intercette di carbonio e azoto: i campi con un effetto casuale positivo per il carbonio tendono ad averne uno positivo anche per l'azoto. Tutte le correlazioni che coinvolgono il fosforo, sia a livello di intercette che di pendenze, sono compatibili con zero, confermando che il fosforo si comporta in modo indipendente dal ciclo carbonio-azoto a livello di campo.

5.4. Selezione delle covariate

La selezione delle covariate è condotta tramite selezione predittiva proiettiva: partendo dal modello completo come riferimento, la sua distribuzione a posteriori viene proiettata su sottoinsiemi crescenti di predittori, scegliendo a ogni passo quello che minimizza la perdita predittiva rispetto al riferimento. Il metodo è *post-hoc* ma privo dell'inflazione della significatività tipica della selezione *stepwise*, poiché non ricerca *pattern* direttamente nei dati ma misura la perdita informativa rispetto al modello completo già stimato.

La Figura 5 mostra il percorso di selezione predittiva per ciascuna risposta. Per carbonio e azoto, Texture_2 emerge come la covariata fisica più rilevante: il suo contributo predittivo marginale giustifica l'inclusione già al primo passo dopo la profondità. Le variabili di gestione non vengono selezionate per nessuna risposta, in accordo con le stime degli effetti fissi riportate nella Sezione 5.1. Per il fosforo, nessuna covariata aggiuntiva migliora la predizione rispetto al solo effetto casuale di campo: il risultato è coerente con l'assenza di effetti *within-field* chiaramente non nulli già osservata nella Tabella 3. Il modello completo è mantenuto come riferimento inferenziale; i risultati della selezione costituiscono una verifica di coerenza interna, non un criterio di riformulazione.

5.5. Confronto tra strutture gerarchiche

Per confrontare i modelli si utilizza la validazione incrociata LOO (Leave-One-Out): ogni osservazione viene rimossa a turno e il modello la predice fuori campione, senza mai averla vista in fase di stima. La qualità predittiva è misurata dall'ELPD (Expected Log Predictive Density): valori più alti indicano migliore capacità predittiva. Il calcolo esatto richiederebbe una nuova stima del modello per ciascuna osservazione rimossa; PSIS (Pareto Smoothed

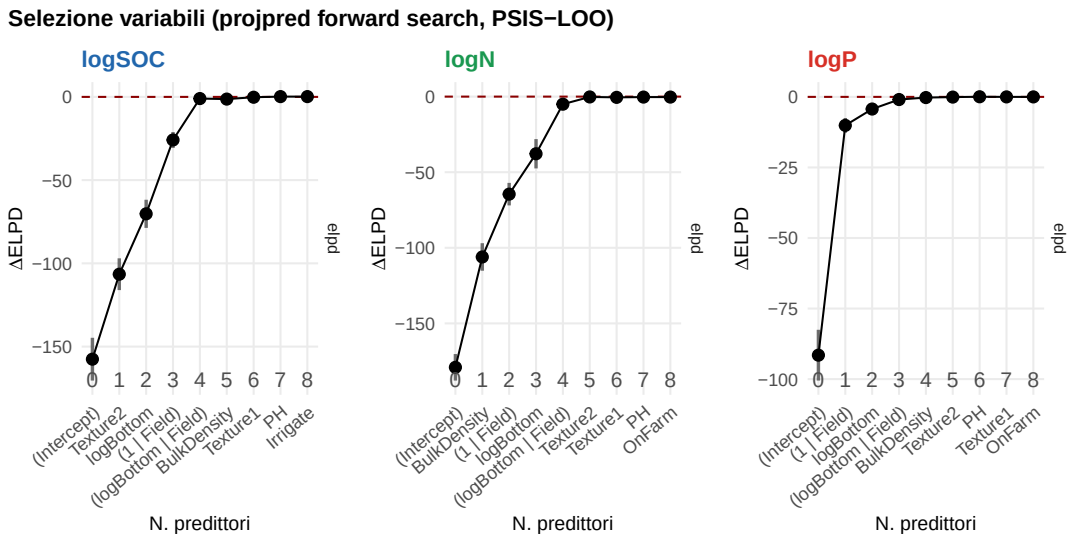


Figura 5: Selezione predittiva per SOC, N e P: ΔELPD in funzione del numero di predittori (linea tratteggiata: soglia di accettazione).

Importance Sampling) lo approssima efficientemente, e l'indice $\hat{k}$ quantifica localmente l'affidabilità dell'approssimazione: osservazioni con $\hat{k} \geq 0.7$ sono monitorate separatamente nell'analisi di robustezza.

Tabella 6: Confronto LOO-CV tra strutture gerarchiche alternative (ΔELPD rispetto al modello finale).

Modello	ΔELPD	SE	$\lvert z \rvert$
Modello multivariato con intercette e pendenze casuali	**0.0**	—	—
Modelli univariati con intercetta e pendenza casuali	−0.4	≈ 4	≈ 0.1
Modello con pendenza di campo vincolata all'intercetta	−4.1	4.2	≈ 1.0
Modello a sole intercette casuali	−9.2	5.7	1.6

La Tabella 6 confronta la capacità predittiva delle diverse specificazioni gerarchiche. Le differenze tra modelli sono modeste rispetto alle rispettive incertezze: il modello finale non presenta un vantaggio predittivo netto rispetto ai modelli univariati con intercetta e pendenza casuale. La progressione verso strutture più semplici porta a deterioramenti crescenti, ma mai chiaramente separati dalla rispettiva incertezza. La preferenza per il modello finale non è quindi motivata dalla *performance* predittiva, ma dalla sua struttura: è l'unico che consente di stimare le correlazioni tra nutrienti a livello di campo, quantità di interesse primario per l'interpretazione nella sezione successiva. Specificazioni più parsimoniose, incluse varianti fattoriali della struttura casuale, non migliorano e talvolta peggiorano la predizione rispetto al modello completo.

5.6. Analisi di robustezza

La robustezza del modello finale è valutata attraverso quattro controlli complementari.

La diagnostica LOO individua 12 osservazioni su 220 con indice di Pareto $k \geq 0.7$, per le quali l'approssimazione *PSIS* è meno affidabile. Stimando il modello sul *dataset* ridotto, le stime dei parametri strutturali (ρ_r, $\tau_{\beta,r}$, γ_r) non mostrano variazioni rilevanti, con una lieve

riduzione attesa della varianza residua $\hat{\sigma}_r$. Le principali conclusioni sono robuste rispetto alla loro presenza.

Sono state confrontate in LOO specificazioni più parsimoniose: una che rimuove le variabili di gestione e Texture_1, l'altra con struttura casuale ridotta per l'azoto. Nessuna delle due migliora la predizione rispetto al modello completo; la variante con struttura ridotta per l'azoto mostra un modesto peggioramento ($\Delta\text{ELPD} = -5.7$, $\text{SE} = 6.0$), coerente con la variabilità sostanziale nelle pendenze riportata nella Tabella 4. Il modello completo è pertanto mantenuto: non vi è guadagno predittivo che giustifichi la riduzione, e per evitare il rischio di selezioni *post-hoc* della struttura casuale.

L'estensione del modello con una struttura di errore multivariata tra le tre risposte mostra correlazioni residue tutte prossime allo zero e un peggioramento predittivo in LOO ($\Delta\text{ELPD} = -3.1$, $\text{SE} = 1.1$). La dipendenza tra i nutrienti è interamente catturata dalla struttura gerarchica a livello di campo, senza necessità di una componente residua aggiuntiva.

L'autocorrelazione verticale rilevata nei profili del carbonio dall'analisi frequentista (§4.1) è assorbita dalla pendenza casuale per campo nel modello finale: la correlazione tra residui di strati contigui è prossima allo zero per tutte e tre le risposte, senza evidenza di dipendenza seriale residua.

Un ultimo controllo riguarda il confondimento spaziale. Le pratiche di gestione e le principali proprietà fisiche del suolo non sono distribuite casualmente nello spazio: `OnFarm` e `Fertilised` mostrano autocorrelazione geografica significativa, `Texture2` e densità apparente (i predittori più rilevanti per carbonio e azoto) la mostrano in misura ancora maggiore. Aggiungendo la longitudine come predittore tra campo, i coefficienti di gestione si spostano sensibilmente, indicando che gestione e posizione geografica non sono separabili nei dati disponibili. Gli intercetti casuali di campo a posteriori non mostrano invece autocorrelazione spaziale residua apprezzabile: la struttura gerarchica, condizionando sulle covariate fisiche spazialmente strutturate, assorbe già la geografia dei nutrienti. Un modello con processo gaussiano sugli intercetti di campo converge senza difficoltà ma con iperparametri scarsamente identificati con $J = 40$ campi e senza guadagno predittivo in LOO. Le stime degli effetti di gestione vanno pertanto interpretate con cautela, in linea con quanto già osservato nella Sezione 5.1.

6. Discussione dei risultati

6.1. Evidenze sulla struttura di correlazione

La correlazione positiva tra intercetta e pendenza casuale del carbonio ($\rho_{\text{SOC}} \approx +0.52$) traduce in termini modellistici l'osservazione esplorativa: i campi con un livello medio di carbonio più elevato presentano profili verticali più uniformi, conservando il carbonio anche negli strati profondi. Questo schema è coerente con la nozione di *stratification ratio* (Franzluebbers, 2002): una distribuzione verticale più uniforme del carbonio segnala una maggiore qualità funzionale del suolo.

Per l'azoto, la variabilità tra campi nella pendenza del profilo esiste ma non è correlata al livello medio del campo: i campi differiscono nella cinetica di decadimento verticale indipendentemente dalla loro dotazione totale. Per il fosforo la variabilità della pendenza è più debole e molto incerta, in linea con la ridotta mobilità di questo elemento nei suoli tropicali.

La correlazione tra le intercette casuali di carbonio e azoto ($\rho_{\text{int}} \approx +0.39$, CI al 90% lontano da zero) è inferiore alla correlazione grezza osservata nell'analisi esplorativa ($r \approx 0.73$): il modello, controllando per profondità e tessitura, separa la quota sistematica legata alla struttura del profilo da quella attribuibile alla covariazione delle covariate fisiche. La correlazione residua riflette il rapporto carbonio/azoto relativamente costante della materia organica stabilizzata (Cleveland and Liptzin, 2007): campi con più carbonio organico tendono proporzionalmente ad avere più azoto organico. La correlazione opera a livello di campo, non di osservazione: dopo aver rimosso l'effetto di campo, le fluttuazioni locali dei due nutrienti alla stessa profondità sono indipendenti. Le correlazioni che coinvolgono il fosforo sono tutte compatibili con zero.

6.2. Evidenze sulle covariate fisiche

Il gradiente di finezza della tessitura ($\texttt{Texture}_2$) è il predittore fisico dominante per carbonio e azoto: suoli con maggiore contenuto relativo di argilla e limo trattengono più materia organica, in accordo con la letteratura sulla stabilizzazione minerale (Matus, 2021). Il contrasto argilla/limo ($\texttt{Texture}_1$) non aggiunge invece informazione predittiva: è il contrasto globale tra frazioni fini e sabbia che conta, non la ripartizione interna tra le due frazioni fini.

La densità apparente mostra effetti negativi chiaramente non nulli per entrambi carbonio e azoto, con un'intensità inferiore rispetto alla tessitura. Per il fosforo nessuna covariata fisica risulta informativa.

Il profilo verticale segue una relazione log-log (legge di potenza), forma funzionale consolidata per i profili di nutrienti in suoli agricoli (Jobbágy and Jackson, 2000).

L'asimmetria tra i due cicli è sistematica e pervasiva: tessitura, densità apparente e profondità governano carbonio e azoto, ma non il fosforo.

6.3. Evidenze sugli effetti gestionali

Due ragioni limitano la possibilità di trarre conclusioni gestionali dai dati disponibili.

Il primo è un confondimento strutturale tra le variabili di gestione, che covariano tra loro: i coefficienti stimati non sono interpretabili come effetti marginali indipendenti. Le stime riflettono un effetto congiunto delle pratiche, non scomponibile nei contributi individuali.

Il secondo, documentato nella Sezione 5.6, è un confondimento geografico: le pratiche agricole sono spazialmente concentrate e non separabili dalla posizione dei campi. In assenza di randomizzazione, gestione e contesto geografico rimangono statisticamente indistinguibili.

Questo è in parziale contrasto con Mutambu et al. (2019), che trovano un effetto positivo di irrigazione e fertilizzazione sui profili di carbonio e azoto in un contesto tanzaniano simile. La spiegazione più probabile è che quel lavoro non controlla per la tessitura del suolo, che nel nostro modello assorbe gran parte della variabilità tra campi e riduce il segnale residuo attribuibile alla gestione.

Più in generale, la letteratura pedologica applicata modella tipicamente i profili con ANOVA stratificata per profondità o con modelli misti a sola intercetta casuale per campo, analizzando i nutrienti separatamente. L'approccio adottato aggiunge due elementi non comuni: la stima esplicita della variabilità inter-campo nella forma del profilo verticale e un *framework* multivariato che consente di quantificare le correlazioni tra nutrienti a livello di campo come quantità inferenziali esplicite.

7. Conclusioni

L'obiettivo dell'analisi era valutare se, e in che misura, le proprietà del suolo varino in funzione della gestione, tenendo conto della struttura di campionamento e della profondità dei prelievi. Il modello gerarchico bayesiano multivariato adottato risponde a questa domanda separando esplicitamente la variabilità tra campi dalla dipendenza verticale. Il risultato principale è che le proprietà del suolo differiscono in misura sostanziale tra campi, ma questa variabilità è spiegata meglio dalle proprietà fisiche del suolo, in primo luogo il gradiente di finezza della tessitura, più che dalle pratiche di gestione.

Gli effetti della gestione non risultano predittivi per nessuna delle tre risposte, e i coefficienti stimati non sono interpretabili come effetti causali, in ragione del doppio confondimento documentato: strutturale tra le variabili binarie, e geografico tra posizione e pratica agricola. In un disegno osservazionale senza randomizzazione, i due contributi non sono separabili.

L'analisi rivela strutture di interesse indipendenti dalla domanda gestionale: la correlazione positiva tra gli effetti casuali di campo per carbonio e azoto, la sistematica indipendenza del fosforo dalle covariate fisiche, e la robustezza di questi risultati rispetto a specificazioni alternative del modello.

Un limite rilevante è l'assenza di interazioni tra profondità e gestione: con i dati disponibili non è possibile verificare se la gestione modifica anche la forma del profilo verticale, oltre al livello medio. Questa rimane una domanda aperta per studi futuri con disegno sperimentale controllato.

Il codice R e Stan utilizzato per l'analisi è disponibile nella repository pubblica: `https://github.com/GabriPaganelli/crop-soil-analysis`.

Riferimenti bibliografici

Cleveland, C. C. and Liptzin, D. (2007). C:N:P stoichiometry in soil: is there a "Redfield ratio" for the microbial biomass? *Biogeochemistry*, 85(3):235–252.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300.

Franzluebbers, A. J. (2002). Soil organic matter stratification ratio as an indicator of soil quality. *Soil and Tillage Research*, 66(2):95–106.

Jobbágy, E. G. and Jackson, R. B. (2000). The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecological Applications*, 10(2):423–436.

Matus, F. J. (2021). Fine silt and clay content is the main factor defining maximal C and N accumulations in soils: a meta-analysis. *Scientific Reports*, 11:6438.

Mutambu, S. L., Mwangi, P., Msanya, B. M., and Kimaro, D. N. (2019). Soil organic carbon stocks under different land use systems in the Kilombero valley, Tanzania. *Advances in Research*, 19(4):1–14.

A. Statistiche descrittive delle variabili nel *dataset*

Tabella 7: Statistiche descrittive ($N = 220$ osservazioni, $J = 40$ campi).

Variabile	Media	SD	Min	Mediana	Max
SOC (%)	1.304	1.047	0.010	1.059	5.489
N totale (%)	0.105	0.062	0.010	0.100	0.390
P totale (%)	0.105	0.098	0.010	0.080	0.920
Profondità (cm)	43.6	17.8	20.0	40.0	80.0
`Texture1` (ILR_1)	−0.635	0.331	−2.040	−0.630	0.458
`Texture2` (ILR_2)	1.118	0.563	−0.412	1.139	2.572
Densità app. ($\mathrm{g/cm^3}$)	1.36	0.35	0.62	1.28	2.06
pH	5.72	0.34	4.84	5.69	6.99
Variabili di gestione (between-field, 0/1):					
`OnFarm`	10/40 campi (25%)				
`Irrigate`	15/40 campi (37%)				
`Fertilised`	20/40 campi (50%)				
`N_Natural`	5/40 campi (12%)				

B. *Trace plot* e *posterior predictive check*

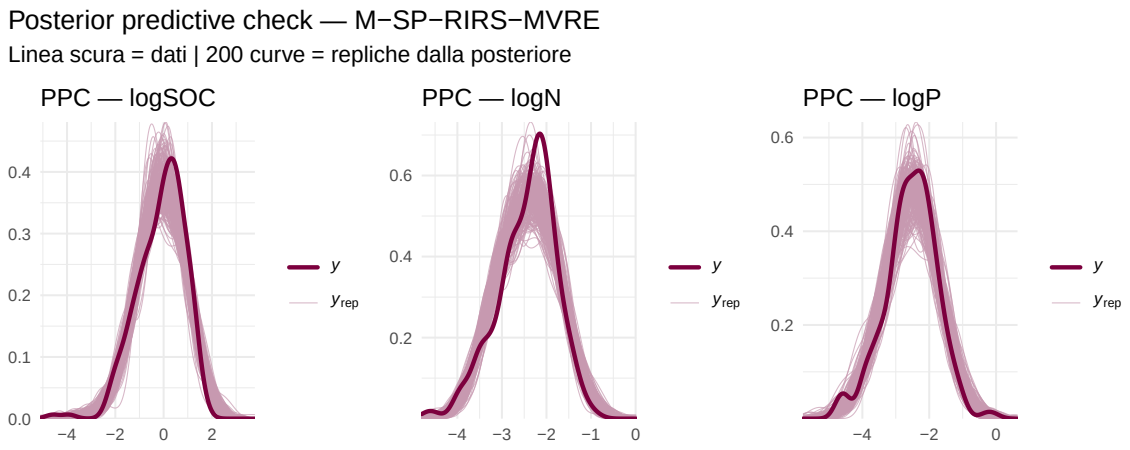


Figura 6: *Posterior predictive check* per SOC, N e P (modello finale).

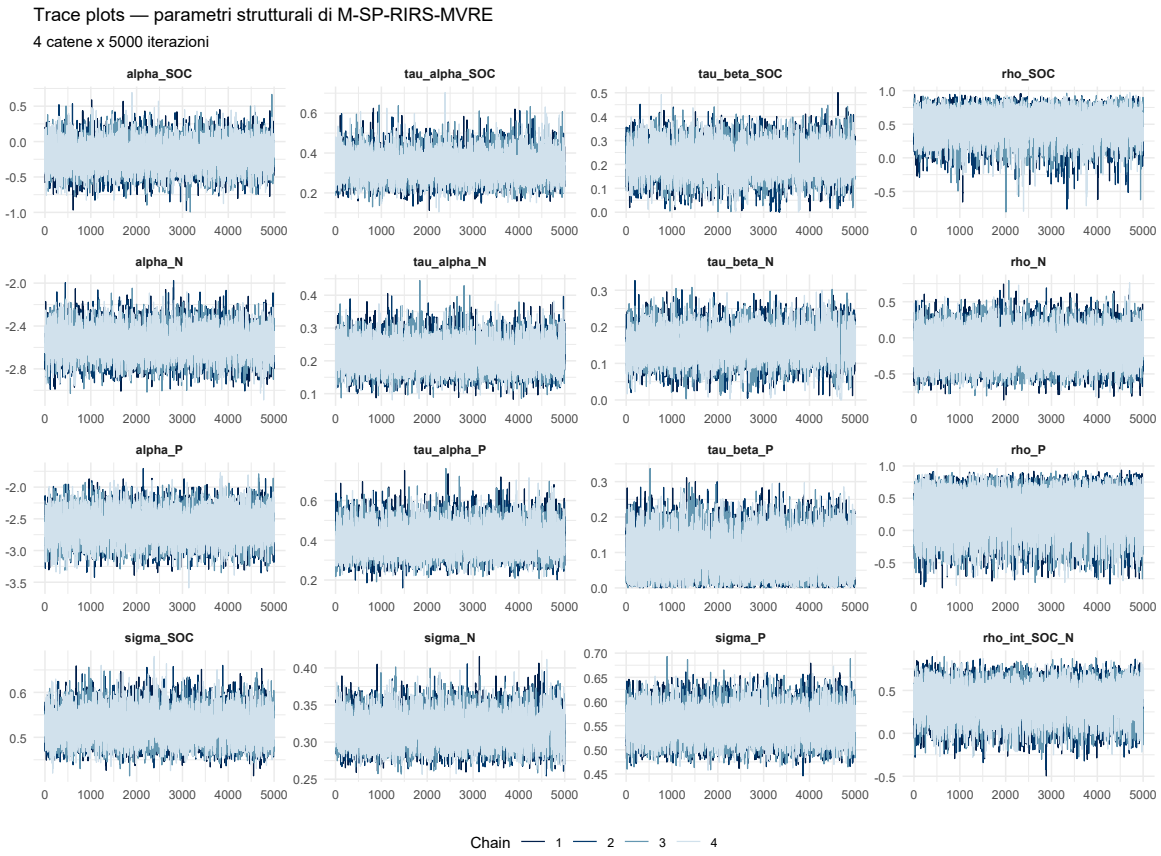


Figura 7: *Trace plot* delle 4 catene MCMC per i principali parametri strutturali del modello finale.